NOVAFCSH
FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

Knowledge Organisation and Terminology: Application to Cork

Margarida Viegas Ramos

PhD in Linguistics
Specialisation in Lexicology, Lexicography and Terminology

November 2020

Thesis submitted to fulfil the requirements for obtaining the doctorate degree in Linguistics

Specialisation in Lexicology, Lexicography and Terminology

And the degree in

Information and Communication Sciences

Developed under the supervision of

Professor Rute Costa

and

Professor Christophe Roche

Funded by the FCT – Fundação para a Ciência e a Tecnologia, Portugal – through the PhD scholarship PD/BD/113972/2015

This thesis has been developed within the scope of a *co-tutelle* agreement between the Universidade NOVA de Lisboa and the Université Savoie Mont Blanc

Tout est signe et tout signe est message

(Marcel Proust)

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my thesis supervisor, Professor Rute Costa, who wisely guided me in this exercise of individual growth. The motivation, along with the multidisciplinary journeys that enriched me as a researcher, would not have been achieved without her vision. There are not enough words to thank her as my mentor, without whom this dissertation would not have been possible, for all the support and constructive feedback.

To Professor Roche, co-supervisor of this thesis, I am thankful for the new world that he presented to me in the context of ontologies and knowledge organisation. Had it not been for the peculiarity of this domain, I would not have crossed the boundaries of the discipline I initially started with. Bridging the humanities and formal logic was both a compelling and a distressful challenge, but the result was undeniably rewarding.

My special thanks to Professor Sylvie Després for the productive and insightful discussions.

I would like to thank Professor Célio Conceição, with whom I took the first steps in Terminology at the UAlg. From the start, the subject has proven to be a promising area of study. Thank you, professor, for showing me the path to the Masters in Terminology at the FCSH-NOVA, but also, for being under the purview of Prof. Rute Costa.

I am also thankful to Professor Frieda Steurs, Professor Sylvie Després and Professor Joana Paulo for the time they dedicated to evaluating my thesis.

A very special thanks to Maria João Ferro for the efficient and accurate review of my English.

I am thankful as well to my fellow doctoral students and doctorates, especially those with whom I worked as a team: Ana, Bruno, Raquel and Sara. Being able to work and collaborate on the team's projects has been a source of continuous learning and I could not be prouder to be part of the LLT group.

I also want to extend a word of gratitude to my colleague Débora, who has been *my person* since our first academic days. Tireless listener, she was always by my side during this journey. I tenderly remember the brainstorming sessions we had.

Last but not least, I am deeply grateful to my family who supported me unconditionally. A special thanks to my life partner, Manuel, who has diligently assisted our children during my absences, and who has always been in good spirits despite the adversities that life can throw at us. To my sons, Ricardo and David, my deepest gratitude for having understood the long moments of absence, even when my body was there but my mind was actually far away. As a final word, I dedicate this thesis to my parents, Celeste and Zacarias, who have always been there for me and who have shown me that it is never too late to accomplish our life goals; but especially to my mother, who has always encouraged me to follow my dreams.

KNOWLEDGE ORGANISATION AND TERMINOLOGY: APPLICATION TO CORK

Margarida Viegas Ramos

ABSTRACT

This PhD thesis aims to prove the relevance of texts within the conceptual strand of terminological work. Our methodology serves to demonstrate how linguists can infer knowledge information from texts and subsequently systematise it, either through semi-formal or formal representations. We mainly focus on the terminological analysis of specialised corpora resorting to semi-automatic tools for text analysis to systematise lexical-semantic relationships observed in specialised discourse context and subsequent modelling of the underlying conceptual system. The ultimate goal of this methodology is to propose a typology that can help lexicographers to write definitions.

Based on the double dimension of Terminology, we hypothesise that text and logic modelling do not go hand in hand since the latter does not directly relate to the former. We highlight that knowledge and language are crucial for knowledge systematisation, albeit keeping in mind that they pertain to different levels of analysis, for they are not isomorphic.

To meet our goals, we resorted to specialised texts produced within the industry of cork. These texts provide us with a test bed made of knowledge-rich data which enable us to demonstrate our deductive mechanisms employing the Aristotelian formula: X=Y+DC through the linguistic and conceptual analysis of the semi-automatically extracted textual data. To explore the corpus, we resorted to text mining strategies where regular expressions play a central role.

The final goal of this study is to create a terminological resource for the cork industry, where two types of resources interlink, namely the CorkCorpus and the OntoCork. TermCork is a project that stems from the organisation of knowledge in the specialised field of cork. For that purpose, a terminological knowledge database is being developed to feed an e-dictionary. This e-dictionary is designed as a multilingual and multimodal product, where several resources, namely linguistic and conceptual ones are paired. OntoCork is a micro domain-ontology where the concepts are enriched with natural language definitions and complemented with images, either annotated with meta-information or enriched with hyperlinks to additional information, such as a lexicographic resource. This type of e-dictionary embodies what we consider a useful terminological tool in the current digital information society: accounting for its main features, along with an electronic format that can be integrated into the Semantic Web due to its interoperability data format. This aspect emphasises its contribution to reduce ambiguity as much as possible and to increase effective communication between experts of the domain, future experts, and language professionals.

KEYWORDS: terminology; domain-ontology; intensional definition; specialised corpus; CorkCorpus; OntoCork; cork

REPRÉSENTATION DES CONNAISSANCES ET TERMINOLOGIE : APPLICATION A L'INDUSTRIE DU LIÈGE

Margarida Viegas Ramos

RÉSUMÉ

Cette thèse vise à prouver la pertinence des textes dans le volet conceptuel du travail terminologique. Notre méthodologie sert à démontrer comment les linguistes peuvent déduire des informations de connaissance à partir de textes et les systématiser par la suite, soit à travers des représentations semi-formelles ou formelles. Nous nous concentrons principalement sur l'analyse terminologique de corpus spécialisé faisant appel à des outils semi-automatiques d'analyse de texte pour systématiser les relations lexico-sémantiques observées dans un contexte de discours spécialisé et la modélisation ultérieure du système conceptuel sous-jacent. L'objectif de cette méthodologie est de proposer une typologie qui peut aider les lexicographes à rédiger des définitions.

Sur la base de la double dimension de la terminologie, nous émettons l'hypothèse que la modélisation textuelle et logique ne va pas de pair puisque cette dernière n'est pas directement liée à la première. Nous soulignons que la connaissance et le langage sont essentiels pour la systématisation des connaissances, tout en gardant à l'esprit qu'ils appartiennent à différents niveaux d'analyse, car ils ne sont pas isomorphes.

Pour atteindre nos objectifs, nous avons eu recours à des textes spécialisés produits dans l'industrie du liège. Ces textes nous fournissent un banc d'essai constitué de données riches en connaissances qui nous permettent de démontrer nos mécanismes déductifs utilisant la formule aristotélicienne : X = Y + DC à travers l'analyse linguistique et conceptuelle des données textuelles extraites semi-automatiquement. Pour l'exploitation du corpus, nous avons recours à des stratégies de text mining où les expressions régulières jouent un rôle central.

Le but de cette étude est de créer une ressource terminologique pour l'industrie du liège, où deux types de ressources sont liés, à savoir le CorkCorpus et l'OntoCork. TermCork est un projet qui découle de l'organisation des connaissances dans le domaine spécialisé du liège. À cette fin, une base de données de connaissances terminologiques est en cours de développement pour alimenter un dictionnaire électronique. Cet e-dictionnaire est conçu comme un produit multilingue et multimodal, où plusieurs ressources, à savoir linguistiques et conceptuelles, sont jumelées. OntoCork est une micro-ontologie de domaine où les concepts sont enrichis de définitions de langage naturel et complétés par des images, annotées avec des méta-informations ou enrichies d'hyperliens vers des informations supplémentaires. Ce type de dictionnaire électronique désigne ce que nous considérons comme un outil terminologique utile dans la société de l'information numérique actuelle : la prise en compte de ses principales caractéristiques, ainsi qu'un format électronique qui peut être intégré dans le Web sémantique en raison de son format de données d'interopérabilité. Cet aspect met l'accent sur sa contribution à réduire autant que possible l'ambiguïté et à accroître l'efficacité de la communication entre les experts du domaine, les futurs experts et les professionnels de la langue. **MOTS-CLÉS :** terminologie ; domaine-ontologie ; définition par intention ; corpus spécialisé ; CorkCorpus ; OntoCork ; liège

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

APCOR: Associação Portuguesa da Cortiça

BNP: Biblioteca Nacional de Portugal

CE Liège: Confédération Européenne du Liège

CL: Corpus Linguistics

COBUILD: Collins-Birmingham University International Language database

CQL: Corpus Query Language

CTKB: corpus-based approach to a terminological knowledge base

DL: Description Logic

ESA: European Space Agency

EUFORGEN: European Forest Genetic Resources Program

FAO: Food and Agriculture Organisation of the United Nations

FSC: Forest Steward Council

GCSE: General Certificate of Secondary Education

GLP: General Language Purposes

ICCSMP: International Code of Cork Stopper Manufacturing Practices

ICNF: Instituto da Conservação da Natureza e das Florestas

IFN6: 6th National Forest Inventory Report

INETI: Instituto Nacional de Engenharia, Tecnologia e Inovação

INPI: Instituto Nacional da Propriedade Industrial

IPQ: Instituto Português da Qualidade

ISO: Organization for International Standardization

ITC: International Trade Statistics

ITU: International Telecommunications Union

KRC: Knowledge Rich Context

KWIC: Key Word In Context

LLELC: Longman / Lancaster English Language Corpus

LM: Lexical Marker

LOD: Linked Open Data

LSP: Language for special purpose

NASA: National Aeronautics and Space Administration

NLP: Natural Language Processing

NP: Norma Portuguesa

OCR: Optical Character Recognition

OWA: Open World Assumption

OWL: Web Ontology Language

PEFC: Programme for the Endorsement of Forest Certification

POS: Part-Of-Speech

R&D: Research and Development

RDF: Resource Description Framework

REGEX: Regular Expressions

SKE: Sketch Engine

SKOS: Simple Knowledge Organization System

SLP: special language purposes

TC: Technical committee

TCA: 2,4,6 - Trichloroanisole

TKB: Terminological Knowledge Base

TW: Terminological Work

W3C: World Wide Web Consortium

XML: Extensible Mark-up Language

TYPOGRAPHIC CONVENTIONS

In this work, some typographic conventions are used in order to differentiate the axis of analysis under focus. With this procedure, terms, concepts, characteristics and conceptual relations are clearly identified and differentiated from each other as exemplified below:

- Terms are written between quotation marks, e.g., “term”
- Concepts are written in two different ways depending on the axis of analysis:
  - between angle brackets and with the first letter capitalised, e.g., `<Concept>; <Concept_1>; <Concept_1_2>` for the linguistic and conceptual analysis; or
  - with CamelBack notation for the ontology representation, e.g., `ConceptExample`
- Characteristics are written between forward slashes, e.g., /characteristic/
- Conceptual relation identifiers are written in italic with an underscore between the forms, e.g., *has_relation*
- Conceptual relations are written with CamelBack notation, e.g., hasRelation

INTRODUCTION

Motivation

This study aims to demonstrate the relevance of texts within the conceptual strand of terminological work (TW).

Our theoretical framework relies on the double dimension of Terminology. In this framework, the linguistic dimension – the term – and the conceptual dimension – the concept – belong to two different but complementary systems since a relationship of mutual interdependence is established between the two. This means that the elements of the knowledge structure are represented though language; thus, texts are at the starting point of our TW.

Based on the double dimension of Terminology, we hypothesise that text and logic modelling do not go hand in hand since the latter does not directly relate to the former. The methodology presented here serves to demonstrate how linguists can infer knowledge information from texts and subsequently systematise it, either through semi-formal or formal representations, and is at the core of this study.

To meet our goals, we resorted to specialised texts produced within the industry of cork. These texts provided us with a test bed of knowledge-rich data, e.g., definitions, which enabled us to demonstrate our deductive mechanisms through the linguistic and conceptual analysis of the terminology – set of terms – that characterises this field of knowledge.

The ultimate goal of this study is to create a terminological resource for the cork industry.

The domain under focus

Cork is said to be the most sustainable material in the world and has been used by man since Antiquity given its unique characteristics.

Currently, the European forestry sector contributes significantly to job creation, with emphasis on the supply generated by the activity of this sector in unpopulated areas, thus contributing to the settlement of populations in those areas. The forestry sector is estimated to employ directly more than 135,000 people (Eurostat, 2010) worldwide, in addition to more than 400,000 forest owners.

In the Portuguese context, the forestry sector is a sector of primary importance, given that it is one of the few sectors whose activity promotes the three main pillars of sustainability: economic, social and environmental.

In economic terms, this sector represents a substantial contribution since it is among others a reliable exporter of tradable goods, and the forestry industries are market leaders in some segments, such as the cork sector. The gross value added of the forestry sector represents 1.2% of the Gross Domestic Product. As it is an economic activity that uses wood and cork as raw material, preferably from the Portuguese forest, the national added value represents more than 70% of the total added value.

As far as the environmental impact is concerned, cork oak forests absorb annually more than 20 million tons of CO^2 (carbon dioxide). Moreover, these forests have the 3rd highest biodiversity in the country and are home to 13,000 species, where the Iberian lynx (only 115 individuals remaining) and the endangered Iberian eagle are included.

The transdisciplinary observed in the domain of cork is at the core of our terminological interest. As we can easily see, cork is a multifaceted field of interest where specialised knowledge and its related terminology – set of terms – is as productive as the wide span of different technological and scientific fields within this vast domain. Besides, Portugal has played a significant role as the leading world producer and transformer of cork for the past few decades.

Cork is a subject of most interest from both a synchronic point of view for the terminological work, and from a diachronic point of view in the perspective of Portuguese cultural heritage.

As far as we could ascertain, much is documented about the domain of cork given its economic, social and environmental impact. However, we have not found evidence of terminological work, namely the organisation of knowledge in the form of a database (TKB) or an ontology[1]. Thus, it is our opinion that a terminological-ontological resource would introduce an innovative aspect to this sphere of interest. This resource is not only designed for experts of the domain, but also for language experts, e.g. terminologists, lexicographers or translators working with complex documents such as standards, where definitions are essential and thus require a common terminology shared by the international community of experts.

Theoretical framework, purpose and methodology

In this study, we highlight that knowledge and language are crucial for knowledge systematisation, albeit keeping in mind that they pertain to different levels of analysis, for they are not isomorphic, i.e., they belong to different semiotic systems. The notion of two dimensions in Terminology entails that for the task of modelling the structure of knowledge we must consider that the resulting model is a multidimensional space where the intersecting axes represent characteristics.

Regarding the framework for knowledge organisation, the standards issued by the technical committee ISO/TC 37[2] are at the core of our theoretical perspective and corresponding terminological choices.

Based on the outlined theoretical perspective, we demonstrate the methodology used to model the results of the two levels of analysis, namely the linguistic and the conceptual analysis in the form of lexical maps for the former and an ontology for the latter, which we named TermCork.

[1] E.g., http://ontologydesignpatterns.org/wiki/Community:Domain

[2] Whose scope of action is the standardization "of descriptions, resources, technologies and services related to terminology, translation, interpreting and other language-based activities in the multilingual information society": https://www.iso.org/committee/48104.html

TermCork is a project that stems from the organisation of knowledge in the specialised field of cork, which is at the core of our study. For that purpose, a terminological knowledge database (TKB) is being developed to feed an e-dictionary. This e-dictionary is designed as a multilingual and multimodal product, where several resources, namely linguistic and conceptual ones are paired to facilitate knowledge acquisition by the user. This type of an e-dictionary embodies what we consider a useful terminological tool in the current digital information society: accounting for its main features, along with an electronic format that can be integrated into the Semantic Web due to its interoperability data format. This aspect emphasises its contribution to reduce ambiguity as much as possible and to increase effective communication between experts of the domain, future experts, and language professionals.

This research mainly focuses on the terminological analysis of specialised corpora resorting to semi-automatic tools for text analysis to systematise lexical-semantic relationships observed in specialised discourse contexts and subsequent modelling of the underlying conceptual system. The ultimate goal of this methodology is to propose a typology that can help lexicographers to write definitions, keeping in mind the different users of dictionaries.

To meet our goals, we have first become familiar with the domain by reading texts produced by experts and semi-experts. Most of the texts produced by the latter have glossaries and explanations of the concepts under focus. Given the large wide span of interests in the domain of cork, we have narrowed down the scope of the study to the subsector of cork stopper manufacturing.

Once familiarised with the domain, we gathered an extensive collection of texts (corpus) and images of the domain of cork and produced a list of terms and definitions of the designated concepts. CorkCorpus is the name of this resource. For the corpus compilation and exploration, we resorted to the Sketch Engine[3] software, which we used to inquire the corpus with text mining strategies. These strategies involve the use of

[3] https://www.sketchengine.eu/

regular expressions in order to extract efficiently from the corpus of analysis specific linguistic expressions generally observed in contextual definitions. Underlying our interest on the recurrence of these linguistic expressions, is the observation of linguistic patterns pointing at lexical-semantic relations between terms, thus providing us with coordinates to interpret the expert's knowledge expressed in texts. These linguistic expressions play the role of linguistic markers since they commonly point at specialised knowledge.

The collections of texts that compose the corpus are multilingual, i.e., these texts are written in Portuguese, French and English. The purpose of these three collections of texts is to observe how concepts are designated in the three different languages. Terms and their corresponding equivalents may be seen in context, a feature that will enable us to (re)write terminological definitions, or to produce glossaries or dictionaries, all based on corpus findings and analysis.

The linguistic analysis of the linguistic markers and corresponding lexical-semantic relations pointed at between two given forms (not exclusively terms) allows us to systematise the interpretation of the information in the form of lexical maps in the first stage of our study. The systematisation of information mainly derives from the interpretation of the lexical-semantical relations pointed at by those linguistic markers, e.g., the relationships of hypernymy-hyponymy and meronymy, but also from the information inferred from the analysis of the term's behaviour in the syntagmatic axis.

After the linguistic analysis, we then step onto the conceptual analysis, which, according to our theoretical perspective of the double dimension of Terminology, is a different level of analysis. Here, despite inferring conceptual relations from the linguistic analysis, we take into consideration a different terminology along with a different modelling representation of knowledge so that the conceptual level is not mixed with the linguistic one. To account for the conceptual level of analysis, where we aim to infer conceptual relations and identify essential characteristics that will later help us through the process of building conceptual maps, we resort to deductive mechanisms employing

the Aristotelian formula: X=Y+DC, to the extent that essential and differential characteristics are deductively obtained and further enunciated systematically.

With the systematisation of conceptual information, we can propose conceptual maps where concepts have long identifiers, i.e., long names. Based on these identifiers and the differential characteristics we have previously identified, we finally propose a model to write intensional definitions, although not only based on the linguistic and the conceptual analysis but also on the knowledge of the domain we had previously acquired.

The two stages of analysis of the definitions extracted from the corpus allowed us to systematise the information in the form of lexical maps in the first stage; to propose conceptual maps in the second stage; and finally, to build an ontology with Protégé[4] – an ontology editor based on Description Logic rules written in Manchester Syntax – in the third stage. In this last stage, concepts are described through a formal language on the one hand and defined in natural language definitions on the other. The terms that designate those concepts are written in Portuguese along with their equivalents in French and English. Underlying the prominence of the Portuguese language is a long history of an active presence of Portuguese experts in the domain and subsequent term coinage.

To build the ontology, we have considered the previously inferred axis of analysis, namely that the physical object designated by "cork stopper" has different states throughout its manufacturing process; different parts; different shapes; different functions; and is made of different types of the same substance, e.g., natural cork vs. cork granules. The task of writing a terminological definition for `<CorkStopper>` in natural language, with the criterion of taking into consideration all those axes of analysis is not an easy one. Furthermore, the outcome of the attempt certainly mirrors an overwhelming definitional context. However, we believe that a well-structured model

[4] https://protege.stanford.edu/

for writing definitions in natural language based on the ontological formal definitions is a plausible solution. A goal that we intend to demonstrate with this study.

OntoCork is a micro domain-ontology where the concepts are enriched with natural language definitions, which in turn are embedded with skos labels[5] – a common data model for sharing and linking knowledge organisation systems via the Web as a W3C[6] recommendation – and complemented with images, either annotated with meta-information or enriched with hyperlinks to additional information, such as a lexicographic resource, e.g., an e-dictionary built with Lexonomy[7] – an open-source platform for writing and publishing dictionaries. Such complementarity of information mirrors our theoretical framework, namely the double dimension of Terminology. Concepts are at the core of the terminological work; however, natural language plays a fundamental role, for terms are the verbal designation of concepts and the means to express knowledge.

One of the standout features we considered when building OntoCork is the possibility of querying the ontology regarding the stage of completion of a given <Cork stopper>, i.e., in what stage of manufacturing the concept classifies according to the operation(s) of <FinishingProcess> it is associated with. With such classification – a feature obtained from the logical rules we have created to formally describe <Cork Stoppers> – we believe that the model of OntoCork can be a valuable instrument to be used in the monitoring of manufacturing processes.

The ultimate goal of our study is to create a multi-functional e-tool which provides a medium for perpetuating the Portuguese cultural heritage in the domain of cork by means of sharing and disseminating the evolution of the domain, both conceptually (in the sense of technology development and consequently new concepts) and linguistically – the terminology shared by a community of experts. According to

[5] https://www.w3.org/TR/skos-reference/

[6] https://www.w3.org/

[7] https://www.lexonomy.eu/

Roche (2005), the conceptualisation of the world and corresponding representation entails the notion of ontology, which today constitutes one of the most promising paths for the modelling, i.e., formally representing, of the knowledge system of terminologies. Hence, with the proposal of such an e-tool, the current knowledge of the domain of cork – where technologies and multiple applications of this unique raw material are continuously innovated – along with the underlying historical evidence of national and international achievements, is perpetuated thanks to the new technologies with one primary goal: the democratisation of its access to a heterogeneous community of users. In our view, such terminological product is a valuable asset for the improvement of international communication purposes, or specialised translation, or even for the creation of multilingual specialised dictionaries within the scope of the special field of cork.

Research questions

A linguist terminologist is usually a non-expert of the domain he/she decides to work with. Thus, the main source of knowledge, i.e., concepts and corresponding terms of the domain, are the texts produced by the experts of the domain.

(i) One of our questions is how can we, as a non-expert, grasp the expert's conceptualisations through the interpretation of texts? As Costa (2006) points out, there are no concepts in texts; instead, texts "talk" about concepts and one of the tasks of the terminologist, if the expert is not readily accessible, is to analyse the expert's choices in a specialised communicative context, i.e., texts produced by and for experts, for they are usually rich regarding terms that designate concepts.

(ii) We have observed in the textual data extracted from the corpus of analysis, where the types of texts are mostly standards and technical texts, that the concept <Colmated stopper> is defined in a footnote in the definition of the <Natural stopper> concept. However, the term "Colmated stopper" is widely used throughout the corpus of analysis. We

question whether the concept denoted by this term should have a definition that is different from the definition of the concept <Natural stopper>.

(iii) To answer the previous question, we ask whether the formal representations of the concepts can give us a critical look at the starting textual definitions. Are representations a complementary asset or merely ancillary?

(iv) Furthermore, and considering that we systematise the knowledge conveyed by texts through the methodology of the division of essential characteristics by specific differentiation from the Aristotelian perspective, we discuss how to represent these characteristics, either as coordinates to guides us through the elaboration of conceptual relations – e.g., associative and partitive – or as a given axis of analysis in the form of concepts, in a higher level of abstraction, such as <Function> or <Parts>.

(v) Finally, we question the sparing relevance of descriptive characteristics mentioned in the literature, for contrastively they help us to model the domain on the one hand, and infer knowledge on the other. Our discussion focuses on the level of their representation in the ontology editor, whether as data property or object property – which in Protégé terminology, the latter corresponds to what we call conceptual relation.

Outline of the thesis

In addition to this introduction, this thesis is divided into six chapters, followed by a conclusion, bibliography and annexes we consider relevant for the purpose of our work.

In chapter 1, we describe the domain on which we have decided to focus our terminological study. We start by presenting the motivation underlying the choice of

working with the domain of cork and the reasons for narrowing down the study to a particular sub-domain. We then dedicate three sub-sections where we make a presentation of the international and national historical outlook to demonstrate the ancient relationship between men and cork. Concerning the current days, a description of the social and economic impact is highlighted, where we present some figures regarding domestic and international trade without going into many details. Finally, a few introductory remarks are put forth regarding the description of the main subsectors of the industry of cork, as well as the issue of quality, which is closely related to the last topic addressed, namely the International Organisation for Standardization (ISO).

Chapter 2 is dedicated to the theoretical framework of corpus linguistics methodology which is the basis for our study. A few definitions of what *corpus* is are outlined as well as an overview of pioneering studies in this area. Corpus linguistics is pointed at as a methodology with a long history in general language lexicographic projects from which a prolific amount of literature is available. The bridge between Terminology and corpora is finally addressed, where a few reference works on which we have inspired our work are highlighted. To conclude this section, the criteria for corpus design that we considered important for our corpus building purposes are highlighted.

Chapter 3 is where we address the practical part of corpus building and it is divided into two main parts. We first describe the purpose and design of the corpus compilation, thoroughly presenting our predefined corpus criteria. From these criteria, the typology of texts, the communicative setting and the language eligibility are highlighted as the essential ones. In the second part of this chapter, we describe the corpus of analysis regarding its composition, management and processing. Corpus processing is addressed in more detail so that we can demonstrate the text mining strategies we have developed to explore the corpus in a more efficient way with *CQL,* a corpus query language where regular expressions (regex) are used to extract specific linguistic patterns. To close this section, we list ten definitions we have extracted resorting to those text mining strategies. Out of those ten definitions, we choose four textual definitions to demonstrate the methodology we have developed to linguistically

analyse texts in the first stage, and based on the results of this first stage, to analyse them conceptually in the second stage.

Chapter 4 is divided into two main sections. The first section is dedicated to the theoretical framework to support our perspective regarding the topic of *definition* and *characteristics*. Great importance is given to Aristotle's work, particularly concerning his theory of definition since the issue of *differentia* is at the core of the discussion. It is in the second part of this section that we demonstrate the linguistic analysis of four definitions of <Cork stopper>. To represent the interpretation of the texts, we resort to lexical maps as an additional support to demonstrate our mechanisms of inference.

In chapter 5, the conceptual analysis of the same four definitions discussed in chapter 4 is addressed. We demonstrate a methodology developed to infer characteristics using the Aristotelian formula X=Y+DC. The set of characteristics obtained with this approach are the coordinates of the axes of analysis to model the domain. Based on these axes of analysis, a set of conceptual relations is designed to assist the domain systematisation in the form of an ontology. Similarly to chapter 4, we resort to conceptual maps as an additional support to demonstrate our proposals for knowledge representation employing specific *differentia* division.

Chapter 6 is where we describe our methods to model the domain of cork with the ontology editor Protégé. The conceptual relations used for the modelling of the domain are based on the characteristics we have inferred and systematised in the previous section. During the description of the ontology's building process, some theoretical aspects are briefly addressed given the high formality of several syntaxes we had to use such as OWL (Web Ontology Language) and Manchester Syntax, which are both closely related to Description Logic.

In the conclusion, we make a proposal of a natural language definition for the concept <Colmated cork stopper> in addition to our final remarks.

Description of the domain

1. Description of the domain: cork

1.1. Motivation

The motivation for working with the terminology – set of terms – and underlying specialised knowledge from the domain of cork is tied with its multidisciplinary sphere of interests, namely its current scientific and technological fields of work, where a large number of R&D (research and development) projects – both private and public[8] – are currently undergoing with an important economic, cultural and ecological impact in the Portuguese context. In brief, although seen as a traditional handicraft trade, the cork sector is currently a niche of work in the Portuguese mainland with continuous exponential growth, given its transdisciplinarity and resulting in a rich terminology in use.

Furthermore, considering that Portugal is the largest producer and exporter of cork in the world, we believe that a terminological study on this domain of knowledge is an excellent source of conceptual information to be systematised in a knowledge database. The aim of this systematisation is the knowledge organisation of the domain under focus to produce a terminological resource for both language specialists – e.g., translators – and experts and future experts, and contribute with a tool where both linguistic and conceptual information complement each other. With this complementarity, we believe that specialised contexts of communication, either in the context of international trade or within the discursive community of experts, acquire a note of quality given the purpose of the terminological work. This means that the aim of the organisation of concepts pertaining to this domain is contributing to a non-equivocal

[8] Such as COMPETE 2020 – within the Management Authority for the Competitiveness and Internationalization Operational Program: https://www.compete2020.gov.pt/noticias/detalhe/inovacao_cortica

communication free of ambiguity as far as possible in both national and international professional contexts.

1.2. The choice of the sub-domain

According to Costa and Pereira (2004), the economy of the cork industry in Portugal depends decisively on two aspects:

(1) the production of cork, as a supplier of industrial raw material; and

(2) the production of natural cork stoppers, as a determining factor, to justify the high costs of the raw material.

Based on these two most relevant aspects of the Portuguese cork industry, and given the vast extent of the cork domain, where several industrial sub-sectors thrive, we have narrowed down our study to one of those two sectors; however, we will provide a brief overview of the main subsectors that comprise the domain of cork (Section 1.8, p. 30).

The primary focus of our study is the cork stopper. Our motivation relates with the fact that the manufacture of cork stoppers is the backbone of this chain of production – i.e., the forestry production of cork oak – since it is the product that holds the most significant share of exports within the scope of the Portuguese agriculture sector. Albeit its light consumption of 30-40% of raw material, yet generating 80% of added value, the cork stopper is the cornerstone of the cork oak chain of production in the national exports – a status still up to date (see INPI 2005).

Cork stoppers are a manufactured product that depends entirely on the domain of cork. Therefore, we will, to some extent, address the super-domain of cork – the source of the raw material – since the typology of cork stoppers is determined depending on the quality of the cork. This means that cork quality – which is conditioned by the high calibre (thickness) of the plank – is a critical factor to determine which final products are obtainable from a given cork plank, right after being stripped from the tree.

The same happens with regards to the intermediate manufacturing processes to transform this raw material, depending on the thickness of the plank. There is one simple goal underlying the determination of the quality and the future of a cork plank: to maximise the use of the extracted cork. Thus, depending on the quality of the cork, one may obtain natural cork products, on the one hand, or products composed of agglomerated cork granules, on the other, where the former require the highest quality and inherent maturity of the tree. In contrast, the latter use cork classified with a lower quality, e.g., leftover pieces of broken planks, lower parts of the tree and cork planks extracted from juvenile trees, just to name a few.

1.3. Cork bark – an ancient raw material

As a raw material, cork has many applications and has been used by man since ancient times. The first known references to the application of cork point to the floating properties of this material. One of the first applications of cork in ancient times was as a floating device, e.g., as buoys in the fishing activity – an application that is believed to have been discovered by the Egyptians in the 4th century BC (see Taber, 2009). Gil (2014) states that cork is a material whose "applications have been known since Antiquity, especially in floating devices and as stoppers for beverages, mainly wine, whose market, from the early twentieth century, had a massive expansion, particularly due to the development of several cork based agglomerates." (p.1). The first references to its applications date back to more than 3000 years BC, namely not only in the floating devices we have already mentioned, but also as a sealant, as material to produce footwear and beehives, or even to insulate houses, as well as applications in household utensils or for therapeutic purposes (see Gil, 2015).

According to Taber (2009), no one knows precisely when someone decided to seal a wine container with cork for the first time. However, it is known from the writings of the Greek historian Thucydides that "the peoples of the Mediterranean began to emerge from barbarism when they learned to cultivate the olive and the vine" (p.8). This

author further mentions that the discovery of pottery circa 6000 BC made it possible for people to store and trade wine – the vast majority of trade in those times relied on only three products: wine, grain, and olives or olive oil. The most popular containers were amphoras and soon were adopted by winemakers, for they could carry a large variety of both dry and liquid products. These amphoras were used for nearly 6000 years and could be found in several sizes. The first pieces of evidence of a kind of stopper to prevent the wine from turning acidic[9] belonged to the Egyptians. By circa 3000 BC, Egypt was the centre of wine production. The methods they used to produce wine are clearly described in frescoes that can still be admired today. However, by that time, cork was not only used in Egypt but also in Babylon and Persia. In addition to its use in fishing gear, cork has also been found in Carthaginian cemeteries in Sardinia in engraved sheets, supposedly used to store precious materials, and also as the lids of urns found in some "nuraghi" – cone-shaped monuments. In some Egyptian sarcophagi, amphoras with cork plugs were also found to store food (see APCOR, 2019).

For economy of space, further notes regarding historical pieces of evidence are available in Annex 1. The inclusion of this topic intends to demonstrate the cultural heritage of this domain.

Notwithstanding, we will highlight in the next lines the pieces of evidence that are closely related to the Portuguese legacy.

1.3.1. Some historical facts in the international context

The systematic exploration of the cork oak trees that characterises the Iberian Peninsula and which still exist today in Catalonia and Portugal only started in the 18th century, when the production of cork stoppers became the primary goal. It was also during this century that the first works on the chemical composition of cork were developed, mostly in studies carried by an Italian chemist named Brugnatelli. The

[9] According to Taber, "winemakers soon learned that air is the enemy of wine. While some air is crucial to get fermentation started and turn the sugar in grape juice into alcohol, the resulting wine will become vinegar if it stays in contact with air" (2009, p. 8).

production of the first compendium on subericulture (the cultivation of Suber family trees) dates back to this period as well. By the end of this century, in 1790, the compendium "Azinheiras, Sovereiras e Carvalhos do Além-Tejo" [holm oaks, cork oaks and oaks of Além-Tejo] was published and signed by a Portuguese author, Joaquim Pedro Sequeira (see APCOR 2019).

In 1700, cork stoppers began to be used and in 1770, with the beginning of the Port wine trade, the cork stopper industry started to flourish in northern Portugal associated with this sector (see APCOR, 2019). Taber (2009) mentions that this was the era when cork stoppers had their most significant boost given the emergence of new bottles – in a more stable shape for both standing on the table and stocking them in stacks – in addition to the signing of the "Methuen Treaty between Portugal and England. This treaty was both a military and a commercial accord that gave privileged trade access to both countries in the other's market" (p.14).

In 1750, the first factory for the manufacturing of stoppers was set up in Girona, Spain, and one hundred years later, the industry was already extended across the country. Finally, in the 19th century, France, Italy and Tunisia decided to join the systematic exploitation of cork oak forests, and countries as different as Russia or the United States also started planting these trees (see APCOR, 2019).

1.3.2. Portuguese cork history in a nutshell

According to APCOR (2019), Portugal was a pioneer regarding environmental legislation, since the first agrarian laws that protect the cork oak forests appeared in the beginning of the 13th century, more precisely in 1209. In 1292, King Denis prohibited the felling of cork oaks in Alcáçovas (Alentejo).

The first reference to cork extraction and use of bark in the tanning of animal skins dates back to the year 1320. Later, in 1438, more references are made to the export of Portuguese cork to Flanders.

During the Portuguese Discoveries (15th - 16th century), the builders of ships and caravels that set out to discover new worlds used cork oak to manufacture the parts that

were more exposed to the weather. They argued that the "sóvaro"[10], as it was called in those times, was the best bonding material for the ships: besides being extremely resistant, it never rotted. Further applications were also found in those centuries, namely in the construction industry: in 1510, several objects made of cork were represented in the window of the chapterhouse of the Convent of Cristo, in Tomar; and in 1560, in two other convents: "Convento dos Capuchos", in Sintra, and the Carmelitas, in Buçaco. These two convents used cork on the walls and ceiling of the cells (see APCOR, 2019).

Further mentioned by APCOR (2019), several initiatives have been launched in recent decades, aiming at research and the design of international standards for the cork industry. The Confédération Européenne du Liège[11] (CE Liège), founded in 1987, stands out: formed by cork federations belonging to several countries, this organisation presented in 1996 the International Code of Cork Stopper Manufacturing Practices, an essential document for quality control in the manufacturing of stoppers.

In the 21st century, cork uses have been spreading, particularly to innovative areas such as Design for Sustainability and Eco-Design. Cork has consistently proven to be a field of interest whose scope of novel applications has been continuously evolving, for new generations of artists seek to create everyday objects from materials that are 100% natural and that contribute to environmental sustainability. Concerning fashion, cork occupies an increasingly prominent place, as well as in other industries, such as transport and sport.

This raw material has been used for many purposes due to its intrinsic properties. It can be found on NASA[12] and ESA[13] shuttles; competition boats; tennis and cricket balls; and even incorporates internationally awarded design pieces. Beyond these exotic

[10] Currently, the name of cork oak is "sobreiro" in Portuguese.

[11] http://www.celiege.eu/

[12] National Aeronautics and Space Administration

[13] European Space Agency

applications, cork is commonly used in the construction industry as acoustic, thermal and vibration insulation (walls, ceilings, floors); false ceilings; wall covering, floors and ceilings; baseboards; linoleum; granules for filling spaces and mixtures with mortars; insulating and expansion or compression joints; as well as for industrial purposes, such as anti-vibration for machinery and insulation for industrial cold (see Gil, 2007). These are just a few of the many applications of cork.

In brief, what stands out in cork is the quality of this excellent raw material and particularly its multiple modern applications and extraordinary ability to meet the current generations' market demands. From footwear industry to pharmaceuticals, and even space shuttle engine components, along with a multiplicity of other applications, we have been witnessing the fast-technological development of the cork industry given the current social trends of ecological awareness and related market requirements.

1.4. The Mediterranean endemic cork oak tree

Cork comes from the cork oak tree, which is known by the scientific community as *Quercus Suber L.*

The cork oak (Quercus Suber L.) is an evergreen broad-leaved tree, from the Fagaceae family, that grows in the forests located in the coastal regions of the western Mediterranean basin. Cork oak trees can live for centuries, between 200-250 years (some authors point at 250-350). They are usually 15-20 meters high, but under ideal conditions, they can reach 25 meters. The stem's diameter at a man breast's height can reach 200 cm. The leaves are 4-7 cm long, dark green on the top and paler underneath, thus forming a round-topped head with a glossy green colour. The acorns are 2-3 cm long with a deep cup. The most interesting characteristic of this tree is its outer bark formed by a continuous layer of suberised cells that constitute the external protection of the stem and branches, which is commonly known as cork – a naturally renewable raw material. The bark can be up to 20 cm thick and corresponds to the dermal system that protects the tree from forest fires (see Boshmonart, 2011; Gil & Varela, 2008).

The principal use of this tree is as a source of raw material for industry, namely its cork, which is obtained by stripping the bark from the trunk. Since the natural goal of the bark is to serve as a measure of protection for the tree during forest fires, it re-sprouts from the stem after the tree suffers any fire damage (see EUFORGEN, 2020). According to Boshmonart (2011), this fire-resistant property has been an evolutionary adaptation to the Mediterranean climate where fire is an important ecological factor.

The cork oak is a tree that thrives in areas exposed to both drought and heavy rainfall. It requires a mild annual temperature and prefers sandy and lightly structured soils. Forest landscapes with cork oaks are biologically diverse, which is why many cork oak forests are protected ecosystems in Europe (see EUFORGEN, 2020).

1.4.1. The cork oak bark

Cork is obtained by stripping the bark from the trunk – a procedure that occurs mainly in summer. The cork bark is manually extracted with the help of specific axes and comes out from the tree in the form of semi-tubular planks, leaving the tree with a thin layer of new cork still covering the functional secondary phloem[14] on the trunk. The first moment of debarking a young tree has to comply with strict forestry guides: it cannot occur before the tree has reached 0.7 meters in perimeter and 1.3 meters high. The cork oak tree cannot be totally stripped from its bark, for it would not thrive if that were to be made (see Gil, 2007).

The first cork harvest[15], from which the so-called virgin cork is obtained, takes place when the tree is approximately 25 years old. It is a cork bark with a very irregular exterior surface. Subsequent harvests can occur every 9-12 years, depending on local legislation. In Portugal, the minimum legal[16] periodicity is nine years (*ibid.*). With the

[14] "Phloem, also called bast, tissues in plants that conduct foods made in the leaves to all other parts of the plant. Phloem is composed of various specialized cells called sieve tubes, companion cells, phloem fibres, and phloem parenchyma cells. Primary phloem is formed by the apical meristems (zones of new cell production) of root and shoot tips." https://www.britannica.com/science/phloem

[15] Called "desbóia" in Portuguese.

[16] Decree-Law No. 155/2004: https://dre.pt/pesquisa/-/search/517471/details/maximized

successive harvestings, the cork tends to develop a more uniform exterior surface. These cork barks are then called *reproductive* or *amadia cork.* The first-time harvested reproductive cork still presents irregularities and is called *secondary cork*[17]*,* therefore, as well as the virgin cork, it is ground in granules and used in the industry of agglomerated cork (*ibid.*).

The first harvest of bark considered suitable for the manufacture of natural cork stoppers – which requires a specific plank calibre – can be carried out after 25 years.

1.4.1.1. The layered structure of cork bark

As described by Taber (2009), the tree has two layers of bark. The inner layer is alive, whereas the outer one has died. Given the successive layer's deaths, the outer bark grows thicker. It is this outer layer that can be harvested every decade without damaging the tree. These outer and inner layers are depicted below in Figure 1.

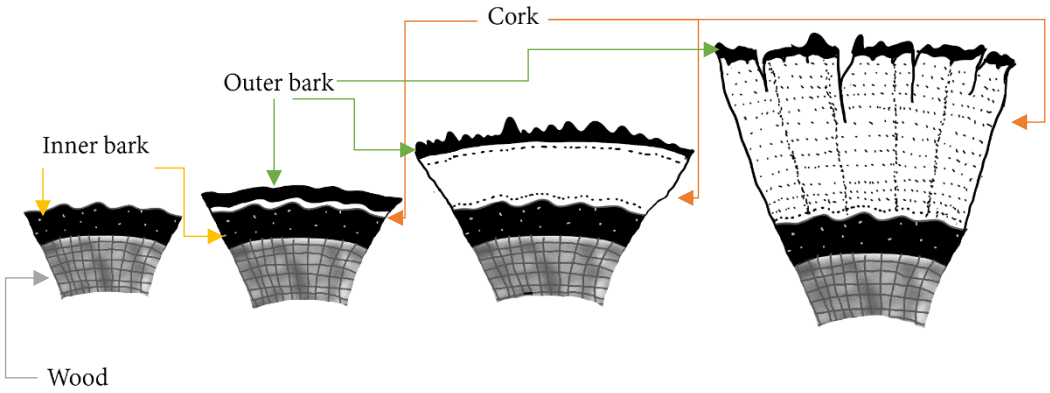


Figure 1: The layered structure of cork bark based on Gil (2007).

Experts call these two layers that constitute the cork oak bark, *meristemic tissues* (Boshmonart, 2011): the *cambium,* which is present in all forest trees that produce *xylem* inside and *phloem* outside, and the cork cambium (phellogen) that generates the

[17] "cortiça secundeira" in Portuguese.

phelloderm inside and the periderm outside. This feature is what underlies the composition of the bark in two parts: the inner layer is called phloem and the outer bark, periderm. The outer bark is not vital to the tree's survival, in contrast with the inner layer; therefore, the former may be periodically withdrawn from the tree without causing any damage, for the inner layer of the tree has the capacity of developing a new outer bark. The purpose of the outer bark is to protect the inner layer – i.e., the living cells of the plant – from the environment. Once the outer bark is stripped, the phellogen (i.e., the cork cambium) dies; however, the development of new phellogen rapidly starts. The tree will respond identically every time the procedure of bark extraction takes place, and that property is the cornerstone of the exploitation of cork (see Boshmonart, 2011).

The following schema represents a cross-section of the cork oak tree trunk, where the several layers described above are systematised:

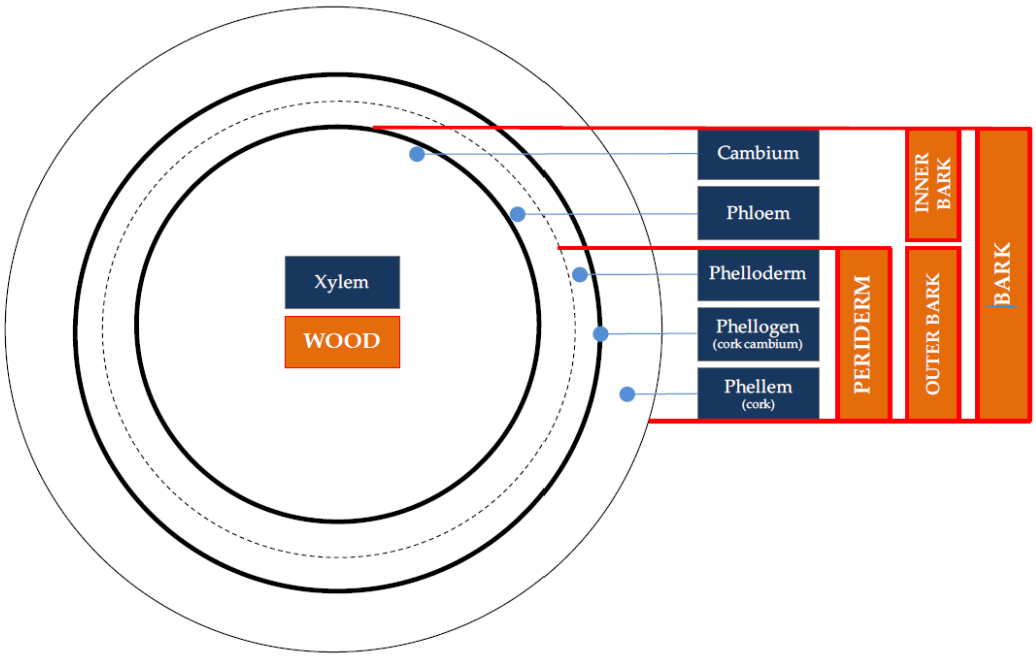


Figure 2: Schema representing a cross-section of the cork oak tree trunk. Source: Boshmonart (2011).

As outlined above, cork is a raw material with unique characteristics: it is 100% natural, versatile and heralded as a sustainable raw material, for it is light, elastic and compressible, impermeable to liquids and gases, with excellent thermal and acoustic

insulation properties, slow combustion, high resistance to friction, hypoallergenic and antistatic. Although it is primarily used to produce wine bottle stoppers, cork is envisaged for a multiplicity of other industrial products as seen above (see EUFORGEN, 2020).

From everything we have said so far, cork is undeniably a multifaceted scope of interest.

1.5. Cork oak forests

According to Boshmonart (2011), most of the current cork oak forests were created in the mid-19th century given the increasing value of cork that derived from the increased demand of cork stoppers. These forests provide multiple economic activities, such as livestock grazing, hunting, and mushroom and honey production. However, the economic value of these activities stems from cork production and its subsequent extraction; if cork extraction were no longer profitable, cork oak forests would be replaced and these other activities might also cease to exist. Besides this significant economic value, cork oak forests provide wildlife habitat, soil erosion prevention and carbon storage, just to name a few of its ecological values.

Given the ecological and socioeconomic value of these forests, it is necessary to adequately manage cork production so it can be guaranteed for the future generations. The quality management of forests is ensured through a certification label of sustainable forest management. Boshmonart (2011) further mentions that this certification is an assurance of quality that indicates that a given product was manufactured complying with an established set of criteria aimed at promoting sustainable forest management. The criteria for such evaluation must describe social, economic and ecological aspects to preserve the forests in the present, as well as in the future. Therefore, several certifications for the label of *sustainable forest management* have been developed. The most extended and widely implemented label is the certification of the Forest Steward

Council[18] (FSC) and the Programme for the Endorsement of Forest Certification[19] (PEFC). In 2011, there were approximately 15,000 ha of cork oak forests in Portugal certified by FSC. This represented 2% of the total surface area; however, at that time, it was estimated by the forestry associations that 150,000 ha would be due in the near future (*ibid.*).

According to the Agriculture, forestry and fishery statistics book (EUROSTAT, 2019), "there are about 182 million ha of forests and other wooded land in the EU, corresponding to around 5 % of the forested area of the world. Forests cover 43 % of the EU's land area" (p.86).

Further advanced by the above report, the

> trees growing in managed forests and removal of logs are main contributors to output from forestry and logging. The main elements of the output from forestry and logging activities are the net increment of forest trees in managed forests, wood in the rough (logs), non-wood products (e.g. cork), and other output (services, secondary activities and other products). In EU forests, trees growing in managed forests and the removal of logs are the main contributors to output from the sector. (EUROSTAT, 2019, p. 91)

Germany produced wood in the rough (logs) with an output value of EUR 4.3 billion in 2016 – one half of total output. While France, Poland and Finland each produced wood in the rough with an output value of between EUR 2.1 billion and EUR 2.8 billion. In that same year, Portugal was the main producer of cork in the European Union. The output value of its non-wood products was EUR 261 million – one fifth (21.4 %) of its total forestry and logging production value (see EUROSTAT, 2019).

1.5.1. The Portuguese forest

According to the IFN6 – 6th National Forest Inventory Report (ICNF, 2019) – Portuguese forest spaces (forest, bush and unproductive land) occupy 6.2 million ha

[18] https://www.fsc.org/en

[19] https://www.pefc.org/

(69.4%) of the national territory. Forest, which includes wooded and temporarily deforested land, is the main use of national land (36%).

Three main species have dominated the Portuguese forest cover since the 1980s: Pinus pinaster, eucalyptus and cork oak. In 2015, the main species in terms of the occupied area were firstly the cork oak (719.9 thousand ha), followed by eucalyptus (845 thousand ha), and finally, Pinus pinaster (713.3 thousand ha) (see ICNF, 2019). These figures are presented in Table 1.

Table 1: Portuguese forest cover by species, based on IFN6 2019

Species 2015	**1000/ha**
Pinus pinaster [Pinheiro-Bravo]	713.3
Eucalyptus [Eucalipto]	845
Cork Oak [Sobreiro]	**719.9**
Holm Oak [Azinheira]	349.4
Stone pine [Pinheiro manso]	193.6
other softwoods [outras resinosas]	52.2
Oak [Carvalho]	81.7
Chesnut [Castanheiro]	48.3
other hardwoods [outras folhosas]	190.2
Carob tree [Alfarrobeira]	16.4
Acacia [Acácia]	8.4
temp. deforested w/o identified species	5.7
Total	3224.1

The number of species in 2015 shown in Table 1 means that in structural, functional and landscape terms, the continent's forest can be organized into four major groups, or forest formations: pine forests (consisting of maritime pine and stone pine); evergreen hardwoods (*montados*, cork oak and holm oak); deciduous hardwoods (oaks, chestnuts and others); and hardwoods for the forestry industry (eucalyptus) (see ICNF, 2019).

1.6. Cork oak landscapes in Portugal: the *montados*

Based on the figures provided by the IFN6 report, the "montados", cork oak and holm oak are the main forest occupation in Portugal, with about 1 million ha and

representing 1/3 of the forest. They are forest ecosystems with multiple uses, whose primary function is not wood production. The 1/3 percentage of forest use places Portugal within the average of the 27 countries of the European Union (see iO10551, 2017).

The cork oak stands are called “montados” in Portugal and are considered traditional multifunctional agriculture, forestry and grazing livestock systems, which prominently characterise the Iberian Peninsula. According to the Food and Agriculture Organisation of the United Nations (2018), these systems

> result from an intentionally induced simplification (both in terms of structure and species diversity) of the Mediterranean forest: anthropic intervention reduces tree density, removes shrub cover (matorral) and fosters the growth of the grass. […] The tree component is oak, usually holm oak (Quercus ilex) and cork oak (Quercus suber), whose acorns provide food for both livestock and wildlife. (FAO, 2018)

Due to its geographical location, Portugal’s mainland is a quintessential cork culture country given the optimal conditions it has for cork production, such as the Mediterranean climate and the soil type, as we can see in Figure 3.

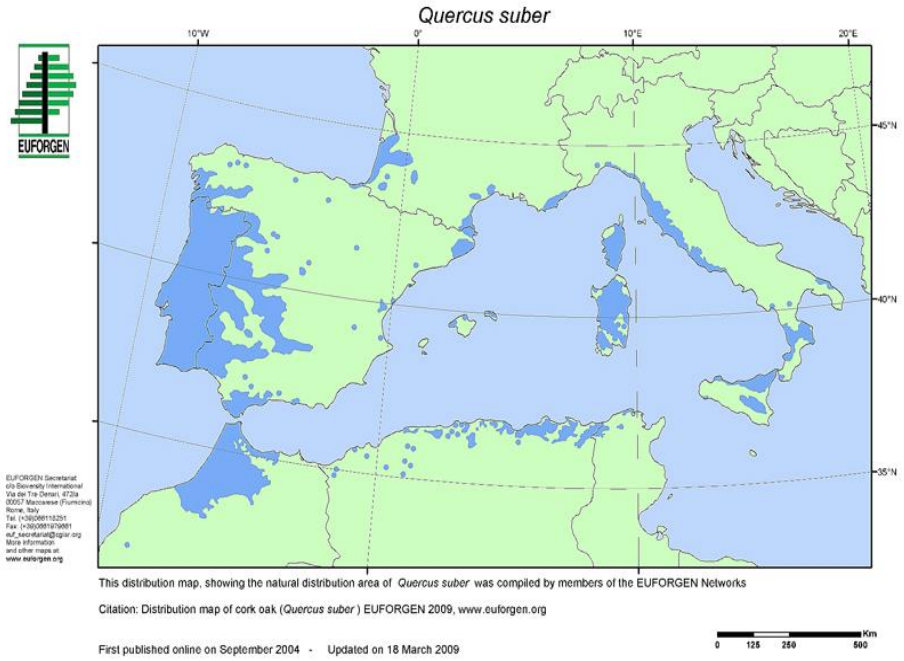


Figure 3: Distribution map of cork oak (Quercus suber) EUFORGEN 2009, www.euforgen.org

According to Gil, the forestry production of cork oaks

> is extremely well-adapted to the semi-arid regions of southern Europe, preventing desertification and providing the perfect habitat for many animal and plant species. Almost the total amount of cork is processed in the European Union, which also imports some cork from Northern Africa, contributing thus to the European economy and employment market. (Gil, 2007, p. 7)

Therefore, the European Union is the largest cork producer – with a share of circa 80% of the worlds' production – and more precisely, the Southern Mediterranean countries, of which Portugal has a significant share of 50% [20] (see Gil, 2007) – this is not a recent phenomenon among the other European and Northern African cork producers. As pointed out by several authors, Portugal has been at the core of cork oak production for centuries (see Gil, 2002; Pereda, 2008).

According to Costa and Pereira (2004), cork production has long been a relevant economic activity for Portugal, with increasing importance since the second half of the 19th century and corresponding to one of the most important and unique products of national export. The industrial transformation of cork developed more slowly, acquiring national importance mainly from the 1960s onwards, when the transformative capacity was decisively developed. At that time, the country started to export mainly finished products to the detriment of exports of raw materials and semi-manufactured products. In fact, until 1960, raw materials represented more than 75% of the total volume of exported cork materials, progressively reducing their weight to about 50% in 1965, 40% in 1975 and more significantly only after the beginning of the 1980s, with 20% of the volume of exports in 1982, a percentage that has remained roughly constant until today.

Currently, approximately 650 cork companies operate in Portugal, employing around 9,000 workers (see FAO, 2018). As one can deduce, cork oak forests have an undeniable economic and social value. The production of cork represents about 0.9% of

[20] According to APCOR (2019), in 2018, the World cork production rose to 201,000 tonnes, with Portugal as the leader in production, with 49.6 %, i.e., 100,000 tonnes.

the national industrial gross added value; 1.2% of gross domestic added value; 2.1% of industrial employment; 2.2% of domestic employment; 9.1% of forestry industry exports in total national exports (see APCOR, 2015; FAO, 2018).

1.7. Cork production – an economic asset

The production of the subsector of the forestry industry and related exports has allowed Portugal to be the leader in the world ranking of international market shares, in the last decades, as shown in Table 2.

Table 2: List of importing markets for a product exported by Portugal - Metadata Product: 45 Cork and articles of cork. Source: International Trade Statistics (ITC)[21]

Unit: Euro thousand

Importers	Exported value in 2015	Exported value in 2016	Exported value in 2017	Exported value in 2018	Exported value in 2019
World	900,780	934,723	987,474	1,148,204	1,063,430
France	161,732	177,808	185,127	209,827	189,275
United States of America	177,608	167,234	168,973	200,041	180,077
Spain	101,057	111,522	132,010	176,996	158,920
Italy	89,573	95,560	99,451	112,040	104,448
Germany	71,534	74,739	73,342	83,150	77,926
United Kingdom	31,179	28,802	31,330	38,794	43,533
Russian Federation	29,732	27,664	28,301	31,244	37,617

As we can observe in Table 2, the leading importers of the Portuguese cork production are France, followed by the USA and finally Spain. The whole list obtained from the International Trade Statistics (ITC) is disclosed in the shape of a map in Annex 2.

[21] ITC calculations based on UN COMTRADE statistics, available at https://www.trademap.org/tradestat/Country_SelProductCountry_TS.aspx

According to the Portuguese General Directorate of Economic Activities (iO10551, 2017), the number of companies that made up the cork subsector in 2015 has remained practically unchanged, standing at 916 entities. Revenue has been increasing consistently, reaching 1.4 billion euros in 2015, mainly due to the manufacture of cork stoppers.

In 2016, the subsector of cork represented 1.76% of total Portuguese exports of goods, a value that has more significance when considering the very positive result of the trade balance of goods (the value of exports is five times higher than that of imports). The main customers are France (19.06%), followed by the United States (17.86%) and Spain (12.03%). The leading suppliers are Spain (74.61%), Morocco (9.02%) and the United States (6.63%). The North of Portugal alone is accountable for 90% of cork exports (see iO10551, 2017).

As shown above, Portugal is currently the largest cork producer and exporter in the world. However, given its cork transformation industry, it is not only an exporter but also an importer of cork. The leading provider of cork is Spain, as seen below.

Table 3: List of supplying markets for a product imported by Portugal; Product: 45 Cork and articles of cork. Source: International Trade Statistics (ITC)

Unit: Euro thousand

Exporters	Imported value in 2015	Imported value in 2016	Imported value in 2017	Imported value in 2018	Imported value in 2019
World	147,324	167,747	175,273	230,732	200,095
Spain	103,611	123,461	134,808	161,728	139,777
Italy	12,575	9,894	13,141	32,480	21,536
Morocco	13,113	15,752	13,373	19,294	18,615
Tunisia	2,867	3,081	2,564	3,649	5,282
Algeria	798	605	1,263	3,810	4,504
United States of America	7,746	11,574	5,115	3,409	4,156
France	2,022	1,100	2,225	2,692	2,251

In Table 3, we can now see the role played by the remaining cork producers. Spain, the country that shares the ideal geographic location for cork oak forestry

production, is the primary provider of cork for the subsector of transformation in the Portuguese cork industry. The subject of subsectors is addressed in the next section.

As mentioned above, in face of this widespread importance of the cork industry and its transdisciplinarity, as well as the inherent terminology that strongly characterizes this field of knowledge, we chose to narrow this study to cork stoppers. That choice, in turn, has taken into consideration the fact that the cork stopper subsector is the core of the forestry production chain and therefore it is the product that holds the largest share of exports within the Portuguese agricultural sector (see AICEP, 2014).

However, we will provide a brief overview of the three main sectors that play an essential role in the cork industry.

1.8. The three subsectors of the industry of cork

The activity of this industry is divided into three subsectors (see Annex 3), which include the activities of

(i) ***preparation,***

(ii) ***transformation***, and

(iii) ***granulation*** *&* ***agglomeration*** of cork products.

The activity of the *preparation of* cork is the first subsector of the cork industry and comprehends several operations to prepare the amadia cork before its commercialisation (see Bicho, 2004). This subsector has to do with **slicing** (traçamento), **stacking** (empilhamento), **boiling** (cozedura), and **stabilising** (estabelização) the cork. However, during the cork bark extraction, an essential activity is also performed, namely the **selection** (selecção) of the cork. This is one of the last stages of forestry production, which in Portuguese is called *subericultura*[22].

[22] After Joaquim Vieira Natividade (1899-1968), the precursor of the scientific identity of subericulture. According to (Pereda, 2008), the text "Subericultura" [1st ed. 1950] was officially praised in the

The *transformation of cork* corresponds to activities that are associated with the manufacture of natural cork stoppers obtained from the activities of "simples talha" [simple carving] or "corte" [cutting] of boiled reproduction cork. According to Gil (2002), this second subsector is very close to the first, namely the preparation, for it makes the link between forestry production and the industry proper, thus, commonly associated to the cork stopper manufacturing. From this association, one can perceive that the boundaries between the two subsectors are not clearly defined.

Finally, the activities of the *agglomeration of cork products* include the production of materials for the construction industry and for the automobile and aeronautical sectors, among others. This stage includes the production and **finishing** (acabamento) of cylindrical batons of granulated cork for the manufacture of agglomerated cork stoppers and component-parts of technical stoppers. The raw material used by this subsector is the waste from the first and second subsectors (see INPI, 2005; Bicho, 2004). According to Bicho (2004), waste cork products are intermediate (in-between) products, for they are products resulting from *transformation.* Some of these products constitute the raw material for the activity of *agglomeration,* in particular for agglomerated cork stoppers, but not exclusively. They are also directly used in the construction industry.

As outlined in the previous paragraph, we have pointed out the reasons why the third subsector of the cork industry is considered the *transforming* one. It does not only use all the waste from the manufacturing of cork stoppers, but also the breakdown of the agglomeration itself and all types of cork with less commercial value or which may not be transformed by carving.

Government Gazette of 30 / XI / 1950, and distinguished as Subericulture Treaty. It is considered, among the forestry community, as "the Book". According to Pereda (2008), Joaquim Vieira Natividade remains today, 58 years after the publication of his best-known work, Subericulture, a focus of attention for foresters and other stakeholders in the world of cork.

With this short introduction to the sector, we hope to have shown the high capacity of (re)usage of cork as a raw material. As stated by Gil (1998), it seems obvious that when it comes to cork, nothing is lost; everything is used.

1.8.1. From the forest to the bottle – a short overview of a natural cork stopper's journey

In the following lines, we describe, without going in too much detail, some procedures and operations that intervene in the process of manufacturing natural cork stoppers, from the first stage of the extraction of cork until the finishing treatment that a cork stopper may undergo. Our aims are (1) to introduce some information in order to mirror the main tasks and stages within the process of producing and transforming the cork as raw material, and (2) to highlight the terms used in discourse by the experts (in the texts we have read to get familiarised with the knowledge of the domain).

The **extraction** of cork is manually performed by **cutting** large rectangular planks and pulling them out from the tree. Currently, the industrial requirements for the cork's calibre – the thickness of the cork – depend on the applications of the raw material, which are mainly directed at the production of stoppers, as the first option. For the manufacture of natural cork stoppers, the extracted cork planks must have a specific thickness, i.e., they have to be more than 27 mm thick. This restriction is the reason why "virgin cork" (cortiça virgem) and "second cork" (cortiça secundeira), whose thickness does not correspond to the required values, are not used for natural cork stoppers (see Pereira, 2007).

Once stripped from the tree, the cork planks are left on the ground and later transported to an adequate place to dry and stabilise. There is need for uniformity and low humidity conditions regarding the ground where the planks will be piled up with the same side (the "belly") facing down. Piles are built as the cork planks arrive from the stand. Furthermore, since no **selection** is made at this stage, cork planks with different thickness and quality are mixed in the same pile due to the cork-quality diversity that exists between trees in the same stand. However, there are some **separate** piles, built with small pieces of cork and "virgin" and "second cork" (see Pereira, 2007). These types

of cork – the 1st and 2nd extraction of a young tree – do not undergo any other triage procedure but this one, immediately after stripping, given their inappropriateness for the manufacture of stoppers (see Gil, 2002).

The stacks of cork planks may be stored in the field or transported directly to the facilities of a factory and wait for *preparation*. The period of storage ranges from a few weeks to a whole year, in order to **stabilise** the planks chemically and structurally (see Pereira). It is at this point that the activity of the preparation proper starts: planks are **staked** in burdens and **boiled** in water. This means that the planks of cork are prepared (sanitised and softened) for the transformation activity. However, before undergoing this operation, the planks need to be **separated** according to their quality, which is determined by visual evaluation of their thickness or defects. Some planks may include low parts of the trunk, named **shocks** (calços); the possibility of soil-derived microbial attacks orders their excision. In the end, between small pieces of cork and low-quality types of cork, this refused raw material is classified as *waste* (refugo) – cork refused for natural cork stoppers manufacturing, yet good enough to be milled by the agglomeration activity and used for other applications (see Bicho, 2004).

Boiling, drying and quality **selection** are procedures that occur several times before the transformation itself (see Gil, 2002). After the **boiling** (cozedura) operation, planks are once again **sorted** into quality grades. The best quality cork is used for the manufacture of natural cork stoppers, while the remainder is used for agglomerates. We must note that sorting is a procedure that involves the actions of **choosing** and **separating**, whose aim is determining the type of activity that each plank is appropriate for, e.g., stoppers, discs or high-quality cork granules. Once the best quality cork planks are **chosen** for the manufacture of natural cork stoppers, they are boiled and stacked once again in order to evenly **stabilise** and flatten their shape, while **drying** the absorbed water.

The production of natural cork stoppers is one of the transformation activities. This type of stopper is obtained from a thick rectangular-shaped piece of cork named

stripe (rabanada) through **punching** (brocagem[23]). Experts explain that stripes are **punched** (brocadas) and to get those stripes (rabanadas) turned into a rectangular shape, the planks had to be previously **sliced** (rabaneadas). At this stage, after being punched from the stripe, the stopper is only a semi-manufactured product, quite far from being a finished product.

A semi-manufactured natural cork stopper undergoes additional operations until it is a finished product. This is where the finishing process plays a role in the transformation activity. Cork stoppers may be sold with a semi-finished or finished status. The client acquires them (a winery, for instance) either unready or ready to be used, depending on the client's purposes or means to finish the stoppers. Briefly, a semi-finished stopper is a stopper that was submitted to any **finishing treatment** (tratamento de acabamento) of the **finishing process** (processo de acabamento), such as **rectifying** (rectificação), **washing** (lavação) and subsequently "drying" (secagem), except the **final treatment** (tratamento final). At this point, the unready-for-use-stopper is either sold (packed and transported) or continues through the finishing process, until it is ready to be used. To be considered a finished product, the stopper must undergo the final treatments, which are **branding** (marcação) and/or surface coating treatment.

1.8.2. The *transformation* subsector

In the following lines, we will only focus on the main product that results from the activity of cork transformation, namely the cork stopper. However, we have systematised the different types of this product depending on the operations that intervene on its manufacturing process: by simple carving or by grinding. This systematisation can be observed below in Figure 4.

[23] Punching is the term of the manual, semi-automatic or automatic process of perforating the strips of cork with a drill (see APCOR, 2010).

Products obtained by simple carving

natural cork stopper

def. [stopper] obtained by punching stripes of cork board (cooked) and subsequent mechanical and / or chemical finishing operations. They are used for various sealing purposes, namely for beverages

natural cork discs

def. [discs] obtained from plank cork by scraping and subsequent pouring. They are used mainly in stoppers [for sparkling wines] and for other purposes

other artefacts

Products obtained by grinding

pure agglomerated

- black thermal agglomerated
- acoustic black agglomerated
- vibrant black agglomerated

granulated

(products obtained from agglutination of granules)

- compound agglomerated
  - floor coverings
  - wall cladding
  - "rubbercork"
  - **agglomerated cork stoppers**

def. [stopper] obtained by cutting the agglomerate rods (formed by extrusion or by moulding in a tube) or by individual moulding. A binder that does not present problems in the contact with food is used in the process. This product can be used as is or in association with discs, as closures

  - rolls
  - other

Figure 4: Systematisation of products from the transformation subsector of cork, based on Gil (1998).

The purpose of the systematisation depicted in Figure 4 aims at inferring the conceptual organisation of the cork stopper subdomain, based on the intervening operations for the processing of cork and subsequent outcomes.

1.8.2.1. The quality of the cork bark after boiling

To briefly address the topic of cork quality and subsequent products that a piece of cork is adequate for given the thickness of the cork bark, we have provided Table 4.

Table 4: Classification of cork: class and calibre, based on Barata and Ganhão (2004) and Gil (2002)

Designation of the cork according to the cork thickness	Lines (‘)	Calibre (mm)	Class	Main products (diameter)
“delgadinha” [very slender] [24]	6’ – 10’	9-22 mm	good	discs (for technical cork stoppers)
“delgada” [slender]	10’ – 12’	22-27 mm	good	discs natural cork stopper (21 mm)
			inferior[25]	colmated cork stopper (21 mm)
“meia-marca” [half-mark]	12’ – 14’	27-32 mm	good	colmated cork stopper (24 mm)
			inferior	colmated cork stopper (24 mm)
“marca” [mark]	14’ – 18’	32-40 mm	good	natural cork stopper (24mm)
			inferior	colmated cork stopper (24 mm)
“grossa” [thick]	18’ – 23’	40-54 mm		cork stopper
“triângulo” [triangle]	> 24’	54 mm		granulated

Incorporated in the activity of *preparation,* the operation of *boiling* (cozedura)[26] has the primary purpose of cleaning the cork. After this operation, which we will not

[24] Literal translation.

[25] Although called “bad” [má] by the experts, we instead opted for *inferior* in English.

[26] According to Gil (1998), boiling includes baling operations, prior transportation / placing in the boiler and boiling proper.

describe in detail, the activity of *choosing* (escolha)[27] is executed upon the evaluation of the (boiled) cork's quality.

The evaluation of cork quality is determined by the calibre and class of the cork bark. While the classification of calibre is normalised by the Portuguese standard (NP-298) – from which we have the designations of *delgadinha* [very slender], *delgada* [slender], *meia-marca* [half-mark], *marca* [mark], *grossa* [thick] and *triângulo* [triangle] – the classification of quality is traditionally made by the *escolhedores* [choosers] based on their empirical knowledge. Here, the designations are *cortiça flor* [flower cork] or *extra* or *superior* and 1st to 7th class.

In brief, the calibre and quality of the cork planks are what determines the final product. In the several stages of the choosing activity, the first separation is made by calibre, followed by the classification by class established by the standards for cork calibre. Usually, cork calibre is measured in lines. Depending on the calibre, the boards have different designations (see Barata & Ganhão, 2004). Finally, after several separations of the cork through *cutting* (corte) or *slicing* (traçamento), performed by the slicer (traçador), the cork boards will designate the sorting and respective applications, in terms of the final product, as shown above in Table 4. Hence, cork boards are chosen by their thickness – calibre – for the production of natural cork stoppers and/or natural cork discs (necessary for the manufacture of certain stoppers), if they do not present critical defects. Some of these defects are excessive porosity or cracks in the *eel-shaped cork* (cortiça enguiada)[28] because these are ways of potential penetration of agents of infection/contamination of the material, a shelter for insects, fixation of dust and soil, just to name a few (see Gil, 2002). Furthermore, and depending on the *good* or *inferior* quality of a given cork plank, it is determined the type of natural cork stopper: *natural*

[27] The cork is selected by specialised workers based on porosity and structural defects. They cut the edges and choose the boards, according to their thickness and quality, after the rudimentary classification carried out in the yards of the factories (see Gil, 1998).

[28] According to the ICNF [Nature and Forest Conservation Institute], the "visual quality of cork" is determined in compliance with 21 types of defects. Available at: http://www2.icnf.pt/portal/florestas/gf/prdflo/mont/qual-defeit

stopper (rolha natural) or *colmated stopper* (rolha colmatada). The latter is an alternative to overcome the non-critical defects[29] of the cork bark.

1.8.3. Cork stoppers – a product from the *transformation* sub-sector

We will focus now on the subsector of *transformation* since the object of our study mainly concerns the manufacturing of the natural cork stopper. From the extraction of cork to the final product, several stages are needed, depending on the type of stopper one wants to produce.

The overall process of the production of cork stoppers is divided into three stages, namely debarking (descortiçamento), manufacturing the stopper (fabrico da rolha), and finishing the stopper (acabamento da rolha), where each stage encompasses different operations as illustrated below:

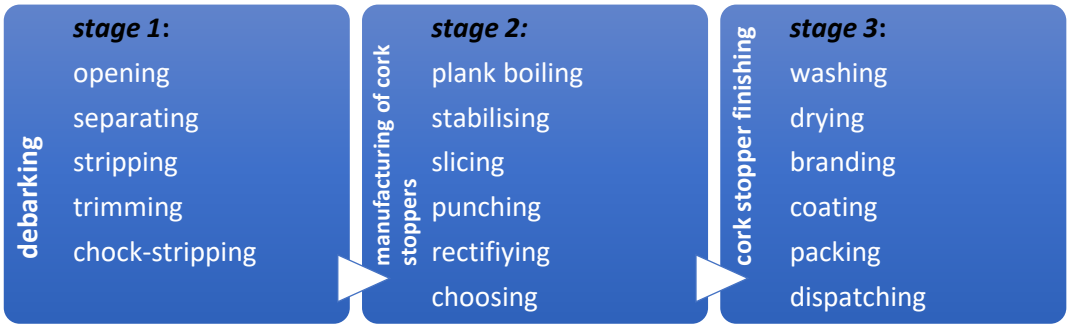


Figure 5: Production of cork stoppers and its different stages, based on Nunes (2013).

The stages of *debarking*[30] and *finishing the stopper* are similar for natural and agglomerated cork stoppers. The manufacturing of cork stoppers (fabrico da rolha) shown above in Figure 5 relates exclusively to natural cork stoppers, while agglomerated cork stoppers undergo other processes that are included in the agglomeration activity, a process that will not be addressed in detail in our study.

[29] The topic of cork and cork stopper defects will not be addressed in this study.

[30] Some authors refer to this procedure as "stripping". However, we will mainly use the term "debarking" when referring to the extraction of the bark from the cork oak tree.

After *choosing* (escolha) the cork planks, an activity that occurs after the operation of *slicing* (traçamento or traçagem) – which are both operations from the subsector of *preparation* – we finally step into the second subsector of the cork industry, namely the main activity of *transforming* the cork, where the cork stopper production is included.

The manufacture of cork stoppers is divided into two production lines, depending on the raw material being used: (1) cork plank or (2) crushed waste from the manufacture of natural cork stoppers. The resulting products of these two lines are natural cork stoppers from the former and agglomerated cork stoppers from the latter.

In the manufacture of natural cork stoppers, and following the terminology previously shown in Table 4, amadia cork planks of *half-mark* calibre are preferably used for the manufacture of the most common dimensions (45x24). This preference is tied with the fact that stoppers are cut perpendicular to the direction of the growth of the cork on the tree, therefore, the calibre of the cork must be greater than the desired diameter for the cork stoppers (see Barata & Ganhão, 2004).

Concerning cork quality, the choice remains at the discretion of the customer and the manufacturer, despite the notion that ideally, only good quality cork planks (1st to 3rd class) should be used for cork stoppers. However, with the constant desire to obtain cheaper cork stoppers, planks of cork with inferior quality (up to 6th) are often used for cork stoppers (*ibid.*).

1.8.3.1. In the line of manufacturing natural cork stoppers

The processing of natural cork stoppers involves a series of operations that will not be addressed at the level of the definition in this section, since the object of study in this work is to organise the typology of cork stoppers, depending on the intervening operations and subsequent terminology used to designate them. Notwithstanding, we have chosen to systematise the operations that intervene in the line of natural cork stoppers manufacturing, as depicted in Figure 6.

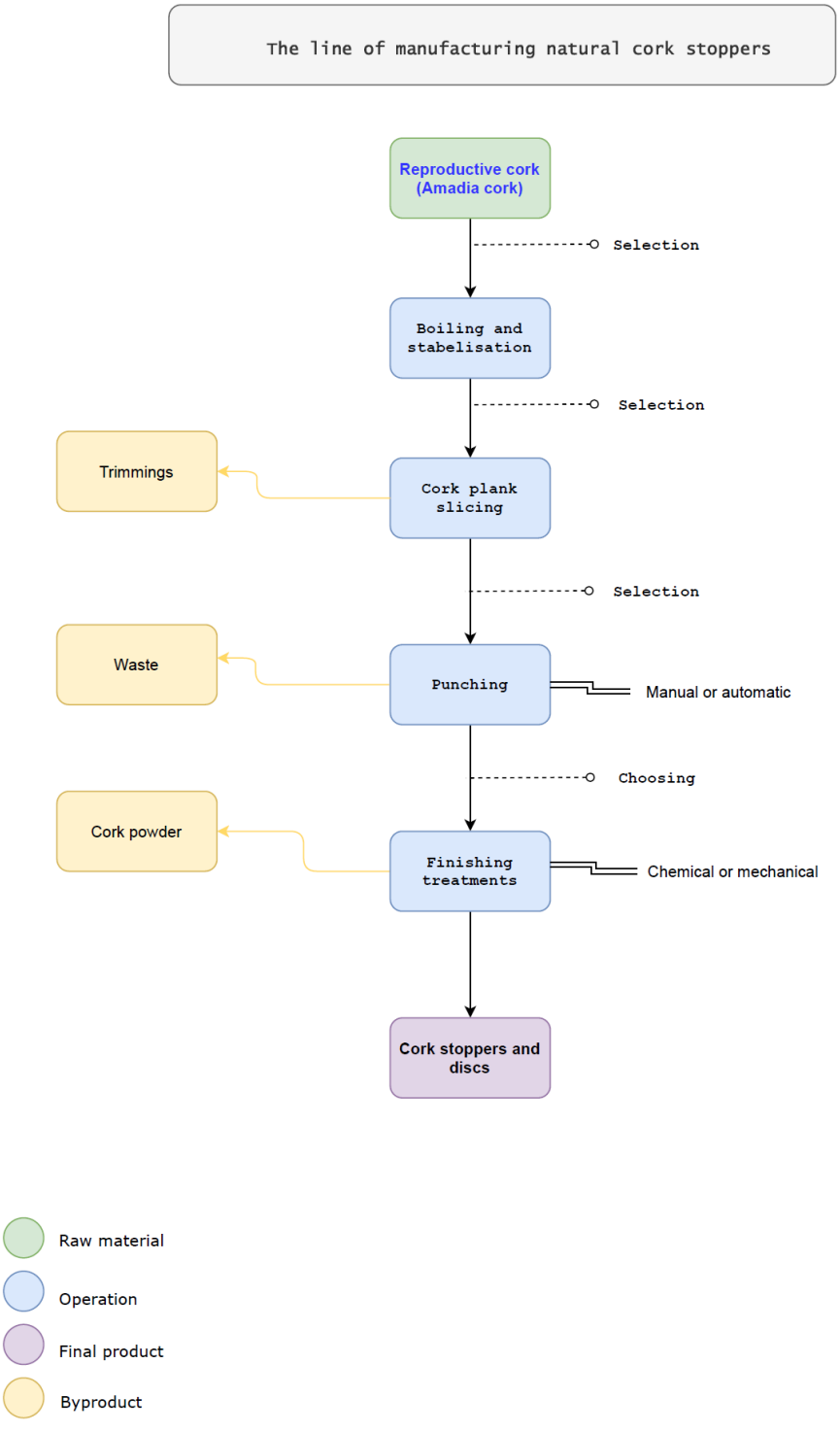


Figure 6: Flowchart 1: The line of manufacturing natural cork stoppers. Based on Gil (2007) and INETI (2001).

We have respected the sequential order of the operations/activities within the manufacturing process in Flowchart 1 (Figure 6). This systematisation does not aim to be exhaustive; instead, it depicts the most relevant operations and corresponding stages where activities are performed, such as "selection", as well as the origin of by-products.

1.8.3.2. In the line of manufacturing agglomerated cork stoppers

Cork is a raw material 100% used and reused like no other material.

As mentioned before, the processing of agglomerated cork stoppers is included in the subsector of agglomeration and makes use of intermediate[31] products (see Section 1.8), i.e., products composed of waste – *dust* (pó); *trims* (aparas); *stoppers with defects* (rolhas defeituosas); *scraps* (bocados), resulting from the natural cork stoppers manufacturing (see Barata & Ganhão, 2004). The granules resulting from the grind of products from waste are called "*clean" granulated* (granulados "limpos") (Bicho, 2004).

The agglomerated cork products are divided into composite agglomerates and pure agglomerates. Composite agglomerates, also called white agglomerates, are the ones used in the manufacture of agglomerated cork stoppers. These agglomerates are composed of cork particles and an adhesive (see Barata & Ganhão, 2004). This adhesive or *binder,* as pointed out by Bicho (2004), must correspond to the type of binders classified as inert, both from the point of view of health or from the aspect of organoleptic[32] changes that they may cause in the food products they have contact with.

Similarly to the previous section, we will not address the intervening operations in the processing of agglomerated cork stoppers at the level of the definition. Here as well, we have systematised them in the form of a flowchart:

[31] Barata and Ganhão (2004) and NP ISO 633 (2011) refer to this raw material as "by-products".

[32] Definition of organoleptic - 1: being, affecting, or relating to qualities (such as taste, color, odor, and feel) of a substance (such as a food or drug) that stimulate the sense organs, e.g., organoleptic research; 2: involving use of the sense organs, e.g., organoleptic evaluation of foods": https://www.merriam-webster.com/dictionary/organoleptic

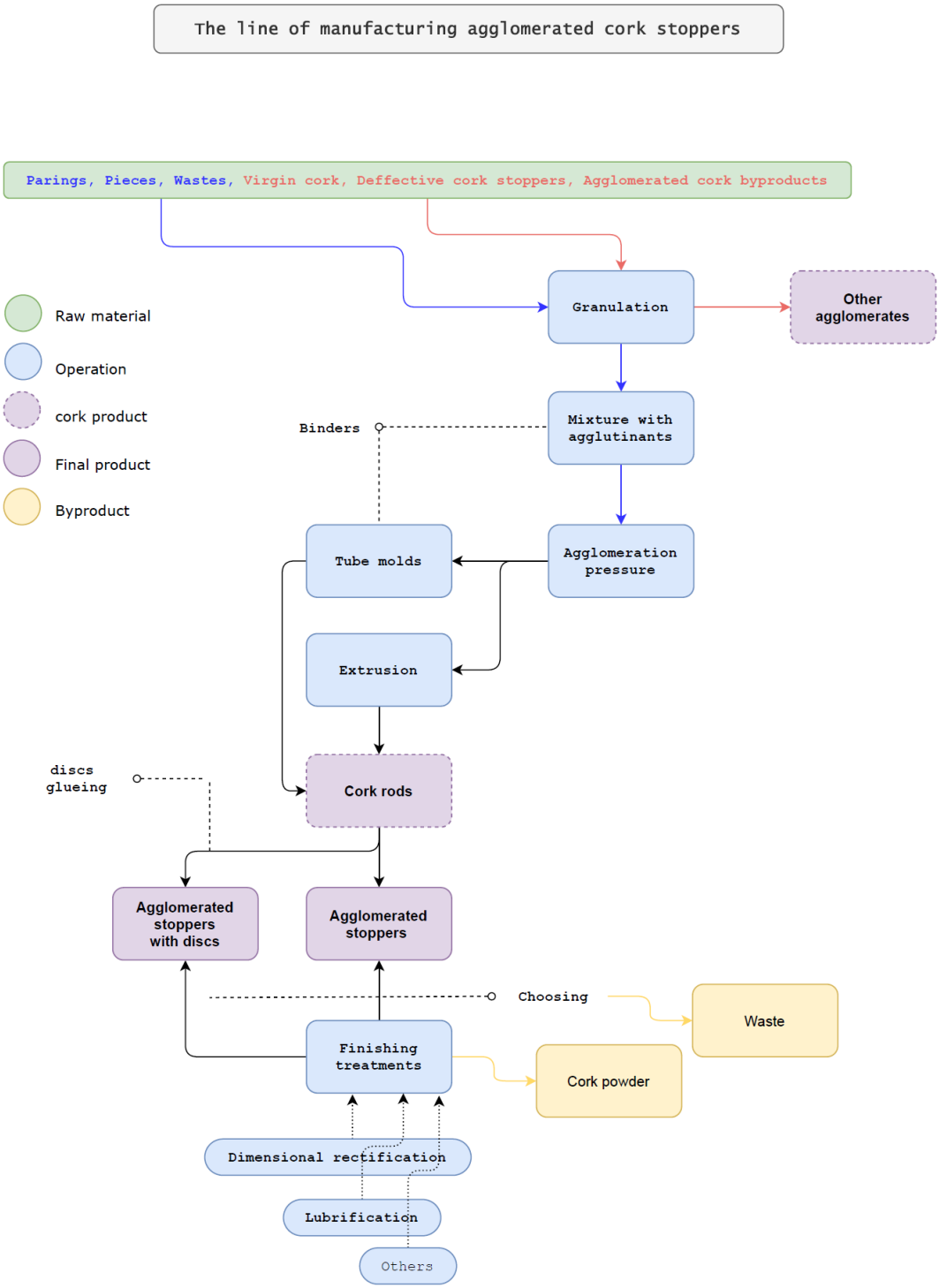


Figure 7: Flowchart 2: The line of manufacturing agglomerated cork stoppers. Based on Gil (2007) and INETI (2001).

Flowchart 2 (Figure 7) depicts the sequential order of the operations and a few (of many) activities, such as “choosing”, within the manufacturing process. Identically to Flowchart 1, it is not meant to be exhaustive; instead, it briefly points at the origin of by-products (i.e., recyclable wastes) and non-valuable cork that will be (re)used as raw material.

Before ending this section, another relevant product must be highlighted, namely the “discos de cortiça natural” [natural cork discs], which is associated with the line of manufacturing agglomerated cork stoppers. As Gil (2002) points out: in the particular context of agglomerated cork stoppers for sparkling wines, since these stoppers have a “body” or “head” made of agglomerated cork, and at the bottom, two or more discs of glued natural cork, this type of stopper encompasses two manufacturing lines: the moulding line and the disc line.

1.8.3.3. Natural cork discs

Natural cork discs are necessary for the manufacture of specific stoppers, namely N + N stopper (rolha N+N) or technical stopper (rolha técnica) (see Norma Mínima V.1, 2007).

As shown in Table 4, Section 1.8.2.1, both *slender* (delgada) and *very slender* (delgadinha) are the types of cork used in the processing of natural cork discs. For this purpose, it is crucial that these corks do not classify as low quality nor present a calibre under 25 mm so that after the slicing of the cork into 80 mm-wide strips, they can be effectively poured (see Gil, 2002).

In the next section, we can finally list the typology of cork stoppers, as a final product from both subsectors of transformation and agglomeration.

1.8.4. Cork stopper typology

In Table 5, we have systematised the types of cork stoppers that are produced (the dimensions are not included, for economy of space).

Table 5: Typology of cork stoppers, based on APCOR (2011)[33] and APCOR (2014)

Type of cork	Designation of the stopper (pt)	Designation of the stopper (en)
cortiça natural [natural cork]	rolha [de cortiça] natural	natural [cork] stopper
	rolha [de cortiça] natural multipeça[s]	multipeace[s] natural [cork] stopper
	rolha natural colmatada { }	natural colmated stopper { }
	rolha capsulada	capsulated stopper
cortiça aglomerada [agglomerated cork]	rolha técnica (1+1; 2+2; 2+0) ~[n+n]	technical stopper (1+1; 2+2; 2+0) ~[n+n]
	rolha de champanhe (0+2; 0+1) { }	champagne stopper (0+2; 0+1) {}
	rolha [de cortiça] aglomerada	agglomerated [cork] stopper
	rolha microgranulada { }	micro-granulated stopper { }

As we can observe in Table 5, cork stoppers are divided into two major types considering the line of manufacturing they result from, i.e., depending on the type of cork used as raw material, as mentioned in Section 1.8.3. Thus, through the different activities intervening in those two lines of stopper manufacturing, there are eight types of cork stoppers in total.

The designations of the types of cork stoppers were obtained from the national standard NP 633: 2011[34] – a text published by the Portuguese Institute for Quality (IPQ)[35] – and are written between square brackets [ ] to indicate they are partially different from the designations put forth by the text APCOR (2014), e.g.:

[33] Document publicly available at www.APCOR.pt.

[34] The standard NP 633: 2011 – "Cork vocabulary" was manually accessed through the Portuguese National Library (BNP).

[35] http://www1.ipq.pt/PT/Pages/Homepage.aspx

NP633:2011	APCOR
natural cork stopper	natural stopper
n+n stopper	1+1 technical stopper

On the other hand, some designations are marked with the empty set { }, when inexistent in the NP 633:2011 standard.

There are other designations introduced by the text Norma Mínima V.1 (2007)[36], however, that do not feature in Table 5, such as "rolha de cortiça aglomerada nova geração" [new generation agglomerated cork stopper] and "rolha técnica de cortiça" [technical cork stopper]. We have decided, instead, to use the terminology of APCOR: "rolha microgranulada" [micro-agglomerated stopper] and "rolha técnica" [technical stopper], respectively.

Those different designations for the same object were the first terminological issue we noticed while familiarising ourselves with the domain under focus, i.e., by reading texts produced by experts. In our view, the existence of multiple designations for the same object introduces ambiguity for a non-expert reader. Nevertheless, our terminological choices between the designations put forth by those two texts will require expert validation in another stage of our project, for they are the ones that master the knowledge of the domain and corresponding terminology.

1.8.5. The quality of cork stoppers

The quality of cork stoppers is determined according to the defects that a stopper might have. In order to assess the quality, the activity *choosing* (escolha) – which is included in the process of stopper manufacturing, right after the operation *punching* (brocagem) (see Flowchart 1, Section 1.8.3.1. p. 39) – serves the manual identification and quantification of the stopper's defects, particularly the ones from the perspective

[36] A document of reference in the subsector, with guidelines and best practices for cork stopper manufacturing.

of its sealing performance, i.e., porosity, structural or manufacturing defects. The identification is achieved through visual observation or optical counting of lenticular channels – the pores of the cork bark – or pneumatic evaluation, based on criteria defined by quality classes (see Gil, 1998).

1.8.5.1. The classification of cork stoppers

Traditionally, the classification of cork stoppers is based on seven visual classes (Bicho, 2004): "superior" or "extra" and from 1st to 6th quality. The selection is performed by comparison of patterns[37] defined either by the factory or the client (see Gil, 1998).

The activities of *choosing* (escolha) and *classifying* (classificação) are both considered two of the most important stages within the scope of stopper manufacturing, thus being critical for the economic performance and the qualitative balance of manufacturing. This means that quality assessment is related to the fact that the price of an extra quality stopper can be several times higher than that of a lower quality stopper (see Bicho, 2004).

1.8.5.2. TCA, the chemical compound 2,4,6 – Trichloroanisole

The chemical compound 2,4,6 - Trichloroanisole (TCA), which is generally expressed in Portuguese as *taste of mould* (gosto a mofo) or *taste of cork* (gosto a rolha) – an expression also known as *cork taint* in English – is a chemical compound, commonly present in Nature, responsible for organoleptic deviation that can be found in cork (see APCOR, 2011).

Currently, there are methods of extracting, preventing and controlling the TCA, harmonized by the ICCSMP rules: International Code of Cork Stopper Manufacturing Practices[38], and others, developed by the companies themselves, which provide them

[37] Visual patterns, according to APCOR (2011).

[38] Document issued by the European Cork federation: http://www.celiege.eu/

with commercial differentiation through the industrial secret of the cork stopper industry. Each company most probably has its variant or process (see Gil, 1998).

1.8.6. Standardisation in the scope of the manufacture of stoppers

Similarly to any industrial branch in developed countries, cork stopper production requires industrial certification in Portugal, whose attestation is the responsibility of the European authority: C. E. Liège - Confédération Européenne du Liège. Designated by the "SYSTECODE" company accreditation system, this certification has the function of attesting that companies work in accordance with the ICCSMP – International Code of Cork Stopper Manufacturing Practices (see CIPR V5, 2006).

In the specific case of the Version 5 of the ICCSMP, the terminology has been updated in accordance with the revision of ISO 633, which summarised the current definitions in other existing standards. Hence, the chain of production has a working tool that is more adapted to the inherent needs of its daily activity (*ibid.*).

1.8.7. ISO: International Organisation for Standardization; ISO / TC87 – Cork

To conclude this chapter, we would like to highlight that Portugal was the forerunner to the normalisation of cork in 1957, through the establishment of the technical committee (CT-16) intending to develop standards for cork and industrial cork products, encompassing raw materials, terminology and finished products (see Gil, 2004).

Within the scope of international standardization, namely in ISO: Organization for International Standardization[39], cork has been under the responsibility of the Technical Committee ISO / TC87 – Cork since 1958, whose Chair and the Secretariat are held by Portugal. Currently, out of the 131 standards (issued or under progress) by this committee, 35 are related to the cork stopper.

[39] https://www.iso.org/home.html

Corpus

2. Corpus

2.1. Corpus definition

The definition of *corpus* is widely discussed in the literature of many branches of Linguistics, such as the field of corpus linguistics, and more recently computational linguistics (see Sinclair, 1991), where the former is described as "a methodology for empirical studies on language" (McEnery & Wilson, 2001; Leech, 2011; Johansson, 2011; Conrad, 2011), and the latter, as the area of "language engineering" (McEnery & Wilson, 2001; Baker, Hardie, & McEnery, 2006).

The following sections outline a few definitions of what a corpus is. We do not intend to provide an exhaustive list but merely a few definitions, and some of them are actually explanations, rather than definitions themselves. Our purpose is to highlight the common characteristics that are advanced by the community of corpus creators, namely terminologists, lexicographers, and language engineers.

2.2. Sinclair's definition

Sinclair (1996) defines *corpus,* not as a collection of texts but rather as a collection of fragments of language that are selected and organised according to explicit linguistic criteria to be used as a sample of the language.

Following Sinclair's work, Tognini-Bonelli (2010) argues that since a corpus is not seen as a collection of texts given the different nature of the fragments of language it contains, it has to be interpreted in a particular manner since the outcomes of corpus exploration are based on excerpts of texts and not on full texts:

> The corpus is not 'just like a text, only more of it'. It brings together many different texts and therefore cannot be identified with a unique and coherent communicative event; the citations in a corpus – expandable from the Key Word in

> Context (KWIC) format to include n number of words – remain fragments of texts and lose out on the integrity of the text. (Tognini-Bonelli, 2010, pp. 19-20)

Moreover, and continuing with Sinclair's theoretical perspectives, Tognini-Bonelli (2010) clearly points at the difference between corpus and text, in particular regarding their purposes:

> the text has a function which is realised in a verbal context, but also extends to a situational and a wider cultural context. It is interpreted by looking at the functions it has as a communicative event. The corpus, on the other hand, does not have a unique function, apart from the one of being a sample of the language gathered for linguistic analysis; the parameters for corpus analysis are above all formal. (Tognini-Bonelli, 2010, pp. 19-20)

Hence, for these two authors, a corpus cannot be seen as a collection of texts, but as *fragments of language* given the different outcomes that each of them can provide. While a text is something one can easily get to know from the beginning to the end, the essence of the corpus, contrarily to that of a text, is not possible to observe directly (see Sinclair, 2004).

2.3. Pearson's choice: McEnery and Wilson's definition

Pearson (1998) outlines several definitions of corpus. Her scientific concern, as a terminologist, was to identify a corpus definition that would adequately support her terminologically-driven corpus criteria.

This author refers to McEnery and Wilson's definition of corpus "as an adequate definition" given the incorporation of the "notions of collection, sampling and representativeness, all of which are important to the description of a corpus" (Pearson, 1998, p. 43). This adequate definition is number three (3), as shown below:

(1) (loosely) any body of text;

(2) (most commonly) a body of machine-readable text;

(3) (more strictly) a finite collection of machine-readable text, sampled to be maximally representative of a language variety.

The above enumeration of several definitions was originally advanced by McEnery and Wilson, 1996 – cited by Pearson (1998, p. 43) – as an attempt of the linguistic community to establish the meaning of what a corpus is back in the 1990s. As Pearson pointed out at that time, the notion of *corpus* was not fully defined by the linguistic community. Yet, later works have dealt with Pearson's concerns.

2.4. McEnery and Wilson's definition

In 2001, McEnery and Wilson developed their notion of corpus under four main touchstones that corpus builders should consider as corpus-design criteria for text collection and subsequent compilation. For these authors, text collection must fulfil the characteristics of (1) sampling and representativeness, gathered in a (2) finite size, (3) machine-readable and (4) a standard reference, bearing in mind that despite the "notion of corpus as the basis for a form of empirical linguistics [it] differs in several fundamental ways from the examination of particular texts." (p. 29). This means that different researchers may approach corpora aiming at different goals, or having different expectations, as long as the corpus itself is not intended to be explored as a text, but rather as a sample of linguistic evidence.

In 2003, McEnery initially described corpus "as a large body of linguistic evidence typically composed of attested language use" and later strengthened the need for a well-organised collection of texts to coherently represent a *sampling frame*:

> The term corpus should properly only be applied to a well-organized collection of data, collected within the boundaries of a sampling frame designed to allow the exploration of a certain linguistic feature (or set of features) via the data collected. A sampling frame is of crucial importance in corpus design. Sampling is inescapable. Unless the object of study is a highly restricted sublanguage […]. (McEnery, 2003, p. 433)

The author's reference to the object of study being a "highly restricted sublanguage" has called our attention.

Many authors refer to *sublanguages* or, synonymously, to *linguistics subsystems*[40] when the discourse of the interlocutors denotes knowledge from a given special subject field, in opposition to communicational contexts where general language is realised. We will assume that the term *sublanguages* is an umbrella term that subsumes specialised discourse.

Our main interest here relies on the possibility of escaping from McEnery's notion of *sampling* when it comes to *sublanguages*. That is, *sampling* is an important criterion for general language usage corpora, but not always necessary or attainable when it comes to restricted ones, such as, for instance, the discourse of experts in specialised contexts of communication. We will address the topic of *sampling* in more detail in Section 2.9.4. (p. 65).

2.5. Baker, *et al.* definition

In 2006, Baker, et al. defined corpus as

> a collection of texts (a 'body' of language) stored in an electronic database. Corpora are usually large bodies of machine-readable text containing thousands or millions of words. A corpus is different from an archive in that often (but not always) the texts have been selected so that they can be said to be representative of a particular language variety or genre, therefore acting as a standard reference. (Baker, Hardie, & McEnery, 2006, pp. 47-48)

This definition is not much different from the ones we have pointed out so far. The interest on this particular definition is to show the four main criteria that characterise corpora commonly referred to in most definitions outlined so far, namely a

[40] The dichotomy *sublanguage* and *linguistics subsystem* will not be developed in our study. Further information can be found in (Sinclair, 2004; Sager, 1990).

(1) collection of texts, (2) machine-readable, (3) representative and (4) standard reference.

2.6. Costa's definition: *specialised corpus*

All definitions outlined up until now are mostly regarding corpora built for lexicographic work, which in practice commonly implicates the analysis and/or description of general language. Hence, the agreement of most authors on the relevance of sample representativeness so that one effectively captures evidences of language from a given community of speakers.

However, in terminological work, the aforementioned types of corpora do not entirely cover the set of terminological purposes, but only partially, since the focus of terminological work – the terminology of a given domain of interest – requires the analysis of *specialised texts,* in the sense of "a stable product resulting from an intellectual and professional activity, coming from a restricted community" (Costa, 2001, p. 60), i.e., texts produced by and for experts of a given field of knowledge.

In 2001, Pavel and Nolet defined corpus as being a "collection of selected written texts assembled for the purpose of performing terminological analysis" (2001, p. 106). This definition points at terminological analysis; however, the type of "selected texts" is not clear – a subject that in our opinion is crucial for corpora compilation when working with terminology.

The definition that clearly identifies the type of texts that should comprise specialised corpora is the one postulated by Costa: "we consider that utterances in specialised contexts [...] constitute a specialised corpus. If the set of specialised statements is representative of the statements produced by the professional class concerned and if the number of statements collected is significant, then we assume that we are dealing with a specialised corpus of reference" (2001, pp. 36-37).

According to this author, the communicative setting and inherent intersubjectivity shared by experts is of utmost importance as a criterion for the eligibility of texts to constitute a specialised corpus within terminological work.

2.7. An overview of pioneering studies in Corpus Linguistics

Working with *corpora* is an empirical[41] methodology with privileged emphasis for the community of language specialists[42] such as lexicographers, language workers, computational linguists, theoretical linguists, applied linguists, among others. The wide array of research areas resorting to corpus analysis pertains to what is called Corpus Linguistics[43] (CL), i.e.,

> […] areas such as language teaching and learning, discourse analysis, literary stylistics, forensic linguistics, pragmatics, speech technology, sociolinguistics and health communication, among others. […] CL has had much to offer other areas by providing a better means of doing things. In this sense, CL is a means to an end rather than an end in itself. (McCarthy & O'Keeffe, 2010, p. 7)

The "means" – working with corpora – is seen, from a methodological point of view, as key for numerous linguistic researches with particular interest on the analysis of language-use in real-life contexts – the *authentic texts*[44] – in contrast to artificial

---

[41] "An empirical approach to knowledge is based on the idea that knowledge comes from our experiences or from observation of the world. In linguistics, empiricism is the idea that the best way to find out about how language works is by analysing real examples of language as it is actually used. Corpus linguistics is therefore a strongly empirical methodology" (Baker, Hardie, & McEnery, 2006, p. 65).

[42] For a thorough explanation of language specialists see (Atkins, Clear, & Ostler, 1992, p. 26).

[43] "Modern corpus linguistics was formed in the context of work on English, though it is now applied to many different languages; it was in this context that techniques such as corpus annotation, and important concepts such as collocation, emerged. Alongside this history of corpus linguistics considered as a methodology stands the history of an alternative approach, sometimes called neo-Firthian, within which the study of words, phraseology and collocation in corpora are the keystone of linguistic theory." (McEnery & Hardie, 2013, p. 1).

[44] According to Williams: "La linguistique de corpus est un domaine qui s'intéresse aux textes, aux textes réels, c'est-à-dire produits pour des raisons de communication entre êtres humains et non des

linguistics productions that are created by traditional linguists to support a given lexical-grammatical theory (see Williams, 2005).

Inspired on the words (above) of McCarthy and O'Keeffe, we must stress that working with corpora is not an end, but rather, a methodology that encompasses a series of tasks in a given linguistic research project involving a large amount of written texts. As mentioned above, it is this large amount of texts, also called as a "collection of texts" – in the words of Francis (1979:110) cit. by Johansson (2011, p. 117) – that commonly stands at the basis of what a corpus is in a very simple way of putting it.

The main stages of and tasks involved in corpora analysis heavily rely on the corpus compilation stage:

(1) a set of well-structured design criteria for corpus building is paramount, in close connection to the purpose of the corpus analysis to be performed, going then

(2) through a laborious work of text capture,

(3) followed by a text typology classification, and finally

(4) the meta-language indexation after the corpus is compiled, among others.

In a few words, corpus-linguists are first and foremost corpus builders – a task that involves sub-tasks – and secondly, interpreters of the corpus evidence.

By following the above steps – which are not exhaustive nor meant to be rigid for the task of building a corpus – it is considered that a note of quality is attributed *a priori* to the corpus evidence. Evidence is observed with the help of natural language processing (NLP) tools, and the quality of the outcomes from the linguistic analysis of the evidence is directly proportional to the predefined criteria and well-structured design for corpus compilation (see Sinclair, 1991).

productions artificielles produites par l'introspection des linguistes, des textes entiers ou du moins des échantillons qui dépassent le stade de la phrase." (2005, p. 13).

Most of the theoretical literature regarding Corpus Linguistics refers to corpus compilation and corpus analysis for the study of general language usage[45]. In our opinion, this literature focus is an outcome of the range of research fields akin to the widely known pioneering projects on corpora building in the 1960s and 1970s. These projects aimed essentially at the study of English[46] language varieties, such as the so-called "Brown family of corpora"[47], from which *comparable corpora*[48] studies were profusely motivated (see McEnery & Hardie, 2013). Identical goals were targeted by some studies developed in the 1980s, such as the well-known lexicographic project COBUILD[49] (Sinclair 1987) – Collins-Birmingham University International Language Database – and the Longman / Lancaster English Language Corpus (LLELC)[50] "both designed for the compilation of English dictionaries aimed at advanced learners" (Laviosa, 2011, p. 132). These last two lexicographic projects are distinct from the Brown Family of Corpora. While the former aimed at comparative research, the latter focused on lexical-grammatical patterns of language (see Conrad, 2011). One of the innovative features of the COBUILD project was its dictionary-making methodology:

---

[45] According to Johansson, "Corpora and the appropriate analysis tools provide an instrument through which we can reveal new things about language structure and use. […]. We have seen a lot already, but much more can be expected if corpora are used with care and imagination. There is probably a bright future for Corpus Linguistics, however it is defined, and – more important – for the study of language in general." (2011, p. 117).

[46] Condamines points out the pioneering studies in English in contrast to French: "Trois grands domaines sont concernés par la description d'une langue à partir de corpus puisque c'est cela qu'il s'agit : la lexicologue (par exemple, Sinclair, 1995), la description de la grammaire, et enfin l'apprentissage d'une langue étrangère. L'utilisation des corpus pour ce type de perspective est nettement plus développée dans la tradition anglo-saxonne que dans la tradition francophone." (Condamines, 2005, p. 40).

[47] The Brown corpus and similar projects: "Brown itself, LOB, and their successors, Frown and FLOB, which sample US and UK English respectively from 1991 rather than 1961 […] thus allowing for diachronic comparison as well as inter-varietal comparison […]. Cross-linguistic comparison is also possible according to the same principle, […] comparison of Chinese and English by contrasting LCMC and FLOB, among other datasets." (McEnery & Hardie, 2013, p. 8).

[48] According to Tognini-Bonelli, "two or more corpora can be designated comparable when they are built on the same design criteria and are of similar size […]. Although the term was first used to designate a variety of multilingual corpora […], corpora which were designed to be compared with each other had already been compiled in the monolingual area" (2010, p. 21).

[49] https://www.collinsdictionary.com/cobuild/

[50] http://global.longmandictionaries.com/longman/corpus

> the dictionary [is] founded on authentic usage in writing and speech. This means that, […], not only is every citation taken from real-life discourse, but the way the different meanings of a word are described and classified can be worked out afresh from the beginning. (Halliday, Teubert, Yalop, & Cermáková, 2004, p. 17)

We must highlight, though, that real-life discourse is not the only focus of analysis for corpus builders and users. Other areas, like register variation were stimulated by the research carried out by Biber, in particular after his work of analysing multiple registers of writing and speech (1988). Another example is the work of Carter and McCarthy (1995) whose focus was conversation features and the differences between conversation and written discourse (see Conrad, 2011).

As pointed out so far, building and using corpora for general language analysis and/or its description in lexicographic resources – just to name a couple of applications – is a broadly used methodology throughout different scientific communities. However, little is said in corpus linguistic literature when it comes to terminological research, unless when translation issues are at the core of the research, as stated by Laviosa (2011): "In the interdisciplinary and international field of Translation Studies, corpora are playing an important role in research, education, professional practice and technology" (p. 143). Moreover, terminology is highlighted by this author as a linguistic source for translators – *terminological data banks / data bases* – but not for corpora-users or corpora-builders.

Above, we have presented a brief outline of the notion of corpus linguistics, its aims, and how multi-purpose this methodology can be. As Conrad (2011) advanced, new researches are constantly arising within the corpus linguistics community:

> Work in these and numerous other fields has continued, but that does not mean that questions in other areas are not appropriate. New research foci are constantly developing, including the incorporation of prosodic analysis in the analysis of lexical bundles (Pickering & Byrd 2008), corpus-based studies of world Englishes (e.g. Nelson 2006; de Klerk 2006) and English as a Lingua Franca (e.g. Prodromou 2008), and

formulaic language use by language learners (Ellis, Simpson-Vlach & Maynard 2008). (Conrad, 2011, p. 53)

Corpus linguistics has been clearly perceived as a useful methodology in a wide range of Linguistics branches in the past decade. At this point, though, one may ask how Terminology[51] – a branch of Linguistics – relates to corpus linguistics.

2.8. Terminology and corpora

In the theoretical literature of corpus linguistics, little is said regarding terminological work. However, many authors have addressed corpus linguistics as a methodology within the scope of Terminology, namely Pearson (1998); Costa (2001); Meyer (2001); Bowker and Pearson (2002); Marshmann (2003); L'Homme (2004); Condamines (2005); Thoiron and Béjoint (2010); Geeraerts (2015), just to name a few. Some of these authors, though, did not consider the task of building a corpus from scratch, where several sub-tasks are undertaken until its accomplishment, such as searching for and collecting specific texts for the compilation of the corpus. Rather, ready-made corpora are suggested and/or were effectively used by these authors as sources of material – sometimes pointed as collections of *running texts* (Pearson, 1998) – for terminological studies.

In Pearson's work, text search and text capture were sourced out from three existing corpora, namely the ITU corpus[52]; the GCSE corpus[53]; and the Nature corpus[54], thus creating three sub-corpora out of those ready-made ones to proceed with her project. The focus of this author's project was the "identification and retrieval of corpus

[51] We differentiate Terminology, graphically capitalised, from terminology. The former designates the discipline, and the latter, the set of terms of a given special field.

[52] International Telecommunications Union – a 4.7m word corpus provided by the University of Edinburgh (see Pearson, 1998).

[53] General Certificate of Secondary Education – a 1m word corpus made available by the Cobuild Unit at the University (*ibid.*).

[54] A 230,000 word corpus provided by the University of Birmingham (*ibid.*).

specific term formation patterns" (Pearson, 1998, p. 123). As far as we could observe, this study involved different types of texts – textbooks; handbooks; and journal articles – and each type corresponded to different special fields of knowledge.

A different approach is observed in (Condamines & Rebeyrolle, 2001), who worked on a corpus composed of a single handbook to pursuit their experiments. The aim was extracting *knowledge rich-contexts* (Meyer, 2001) for the construction of a CTKB (a corpus-based approach to a terminological knowledge base). In this case, the study focused on one special field of knowledge, and the corpus for analysis was a ready-made corpus.

The novelty of creating a *specialised corpus* from scratch was advanced by Costa (2001). We are able to observe in this author's work, a domain-specific corpus built from scratch and composed uniquely of *specialised texts,* i.e., texts produced in a specialised communicative setting, where the experts' discourse is denoted by their linguistic choices while communicating within their community of expertise. In this work, the subject of specialised texts and specialised discourse is thoroughly debated. The domain of the corpus under analysis was *remote sensing*, thus a high level of technical discourse served Costa's observations: from the technicality observed on the typology of texts used to build the corpus, Costa advocates that a corpus built with specialised texts is what defines a *specialised corpus*, as mentioned before.

Last but not least, Marshmann (2003) also embraced the task of building a corpus from scratch in her terminological work. This author searched for and captured specialised texts from the internet, as well as from term banks, with a specific goal: to compile a "corpus spécialisé". The compilation of this specialised corpus was accomplished given the collection of texts produced from different special fields of knowledge: "l'informatique, le droit, la mécanique, et la médicine" (p. 12). Thus, instead of being a specialised corpus with a well-defined domain, Marsham's corpus had several special fields of knowledge to work with.

These are just a few of the authors that have been doing terminology-driven corpus work; many others could be listed here. As mentioned by L'Homme (2004) and

Agbago and Barrière (2005), the number of manually made corpora for terminological purposes is as large as the number of each terminological work because each terminological work requires a new corpus.

There is a keynote shared by the above terminological studies, and in our opinion it is the cornerstone of the semasiological approach of the terminological work when it comes to knowledge extraction from written texts: all these authors highlight the need to work with specialised texts, which is seen as a paramount criterion for the terminological work for the simple reason that they convey domain-specific knowledge (see L'Homme, 2004).

Our research is heavily inspired on these authors' work, not regarding their research goals but the explicit criteria for corpus building, bearing in mind that criteria are motivated by the methodology that underlies corpus-based or corpus-driven researches.

2.9. Criteria for corpus design

The *corpus design* or *linguistic design* (Atkins, Clear, & Ostler, 1992) is the first of many stages in corpus compilation, and it is highlighted in the literature as the most important step prior to text capture. The design is what establishes *a priori* what type of corpus is going to be constructed, thus, the first task of corpora-builders is to seek to obtain answers via a list of questions they should previously elaborate, in order to create a set of criteria that will fulfil the purposes of the corpus:

- is it a diachronic or a synchronic corpus? The former serves, for instance, the observation of the evolution of a given language, while the latter, the observation of a given period of time;
- is it a closed or an open corpus? Where the former restricts the number of texts, thus the need of the criterion of a *numerus clausus* of text samples;

- which types of texts to assemble?
- and what about the range of the language variety?

These are some questions that address some of the criteria used for corpus design. The point is, designing a corpus requires that corpus-builders carefully think about the corpus purposes (see Atkins, Clear, & Ostler, 1992; Baker, Hardie, & McEnery, 2006). Therefore, the purpose of the corpus is fundamental to stipulate its design criteria. The tasks of searching for and capturing texts are dictated by the predefined criteria.

It is important to stress though that some of those criteria are not mandatory, nor finite. Each research has different goals, thus different corpus designs. In fact, there "are now so many corpora for so many purposes that it is impossible to list them, and only a sketchy classification can be attempted." (Tognini-Bonelli, 2010, p. 20). Hence, the results of the corpus analysis will be then just as good as the well-structured set of criteria used to build it (see Sinclair, 1991).

2.9.1. Four main criteria for corpus building

According to the perspectives of both the communities of terminologists and corpus linguists, the main criteria for corpus design are:

(1) machine-readable

(2) finite or non-finite size

(3) sampling/representativeness

(4) balance

Each of these criteria are discussed on the following sub sections.

Instead of *standard reference*[55] (a criterion pointed by McEnery and Wilson, see Section 1.3), we have opted for *balance*. The criterion *Standard reference* is intended to

[55] According to Baker, et al: "The term 'reference corpus' may also be used to describe any corpus that, like [the Brown family] corpora, is not a sample of any particular language variety, domain or text type,

stand for a sampling frame of general language, which is not the focus of our study; therefore, we have not considered it necessary. Balance, though, is pointed as a fundamental criterion, in the theoretical perspectives of both communities of terminologists and corpus linguists, as discussed below.

2.9.2. Machine-readable

The notion of corpus is currently associated with the term *machine-readable* because of its electronic nature: it is a *body* of language material which exists in electronic form, "and which may be processed by computer for various purposes such as linguistic research and language engineering" (Leech, 2013, p. 1).

According to Leech (2013), the 1980s were very prolific regarding corpora building and subsequent development of natural language processing (NLP) tools to explore them. Large corpora came to light in a variety of sizes as an outcome from the high capacity of modern computers. With these computers, searching, processing, and storing texts increased in such a way that "an increasing range of software [was] developed to process corpora and access the information they contain" (*ibid.*), such as concordancers[56], taggers[57], and text analysis software[58]. These tools for corpus analysis are developed in areas of research within computational linguistics, which is why some researchers refer to corpora as a test bed for their work in language engineering; while others refer to them as a repository of language attestation for lexicological purposes, e.g., dictionary-making. In other words, while the former develop corpus tools and/or

but is instead an attempt to represent the general nature of the language through a wide-sampling corpus design." (2006, pp. 136-137).

[56] Also known as a "Key Word In Context (KWIC) concordancing program, which produces displays [of] all occurrences of the word of interest […] lined up beneath one another, with surrounding context shown on both sides." (Manning & Schütze, 1999).

[57] is a "software which automatically carries out tagging on a corpus" (Baker, Hardie, & McEnery, 2006, p. 153) like, for instance, the task of "automatic part-of-speech tagging [that] can be carried out on a corpus, whereby every word within it is assigned a particular grammatical tag" (*ibid.*) (p. 67).

[58] For example, the tool Sketch Engine (https://www.sketchengine.eu/).

corpus enrichment, the latter use the developed tools for the analysis of corpora, whether enriched or not.

Machine-readable corpora have several advantages over the original written format in hard copies. The first and most important advantage of machine-readable corpora

> [...] is that they may be searched and manipulated in ways which are simply not possible with the other formats. [...] With a machine-readable corpus, the [text analysis] task may be accomplished in a few minutes using concordancing software [...]. The second advantage of machine-readable corpora is that they can be [...] easily enriched with additional information. (McEnery & Wilson, 2001, pp. 31-32)

The additional information mentioned by these authors is what helps the linguist to explore the annotated corpus in a faster and more effective way when compared to a raw corpus. The added value of the annotation, together with appropriate tools for corpus analysis, offers the linguist a wide range of approaches to the corpus. The linguist can extract data (language evidence) from the corpus by means of specific queries, where lemma[59], Part-of-Speech (POS), or morphosyntactic structures are parametrised with the help of artificial languages used in computer science, such as *regular expressions,* also known as *REGEX* – a feature commonly used in corpus query languages (CQL) with natural language processing (NLP) tools, such as Sketch Engine (SKE). The subject of REGEX and CQL will be further developed in Section 3.4.1 (p. 103).

The enrichment of corpora with additional information is what is called corpus *annotation*[60]. The outcome of this process of annotation is an *annotated corpus*. On the

[59] "is the base form under which the word is entered and assigned its place: typically, the 'stem'; or simplest form (singular noun, present/infinitive verb, etc.)" (Halliday, Teubert, Yalop, & Cermáková, 2004, p. 6).

[60] According to Leech, "Annotation can be thought of a kind of 'value added' to the raw form of the corpus. Each level of annotation (POS tagging, parsing, semantic tagging, discourse annotation, etc.) adds additional information about the linguistic form and content of the text, and therefore enables us to retrieve from the corpus instances of the phenomena so represented. In this way, the searching of the corpus, or extraction of statistical data from the corpus, can be made more powerful and abstract." (2011, p. 168).

contrary, a corpus that is not processed with any kind of analytic annotation – e.g., POS tagging[61] and lemmatisation[62], among others[63] – is called a *raw corpus* or an unannotated corpus.

We will not address the subject of annotation in much more detail in our study, although some general definitions will be approached in Section 3 (p. 74).

2.9.3. Size

The literature on corpus linguistics does not describe a *finite size* as a fundamental criterion to consider for corpus design. Instead, the corpus-builder must predetermine if the corpus will be either closed (also called static) or open.

The criterion of finitude is what defines a static corpus – a sample of texts that is intended to be of a particular size. Once the target-size is reached, no more texts are included in it. Most static corpora provide a *snapshot* of a particular language variety at a given time (Baker, Hardie, & McEnery, 2006).

On the contrary, open corpora are designed to be dynamic, composed by an open-ended collection of texts (see Atkins, Clear, & Ostler, 1992; McEnery & Wilson, 2001; Bowker & Pearson, 2002; Laviosa, 2011; Johansson, 2011), and have a specific purpose:

> a dynamic corpus is one which is continually growing over time […]. Dynamic corpora are useful in that they provide the means to monitor language change over time – for

[61] this is a "type of annotation or tagging whereby grammatical categories are assigned to words (or in some cases morphemes or phrases), usually via an automatic tagger although human post-editing may take place as a final stage." (Baker, Hardie, & McEnery, 2006, p. 128).

[62] is a "form of automatic annotation that is closely allied to the identification of parts-of-speech and involves the reduction of the words in a corpus to their respective lexemes. Lemmatisation allows the researcher to extract and examine all the variants of a particular lexeme without having to input all the possible variants, and to produce frequency and distribution information for the lexeme." (Baker, Hardie, & McEnery, 2006, pp. 103-104).

[63] The various kinds of corpus annotation and encoding are "orthographic representation, textual and extratextual information, part-of-speech tagging, parsing, semantic annotation, anaphoric annotation, phonetic and prosodic transcription and problem-oriented tagging" (McEnery & Wilson, 2001, p. 73).

> this reason they are sometimes referred to as monitor corpora. (Baker, Hardie, & McEnery, 2006, p. 64)

This dynamic functionality of an open corpus is also emphasised by Bowker and Pearson (2002), although with different research purposes. Instead of corpora for general language purposes (GLP), these authors worked with corpora for special language purposes (SLP)[64]. For these authors, an *open* or *monitor corpus*

> is a more flexible entity to which you can add and remove texts to reflect the changing state of language. Specialized language is typically dynamic — concepts in specialized subject fields are constantly evolving and the terms used to describe these concepts also change. […] Given the dynamic nature of specialized language, an open corpus that can be updated on a regular basis is likely to be more appropriate for many of your LSP needs. (Bowker & Pearson, 2002, p. 48)

Pearson (1998), however, did not consider this criterion of open corpora for the corpus design although the purpose of her project aimed at the terminology of two special fields.

As pointed out so far, there are no definitive guidelines for corpus design criteria. What is perceived, though, is that specific criteria are "determined by your needs and by the goals of your project" (Bowker & Pearson, 2002, p. 45). The notion of an open corpus, and the aspect of continuously feeding it with new texts has inspired our work, as further mentioned in Section 3.3. (p. 87).

2.9.4. Sampling

A corpus is considered to be a sample of a language or language variety, the latter corresponding to the language used by the population that the corpus intends to represent (see Laviosa, 2011). However, representing a given language is a quality that corpora are unlikely to accomplish, as thoroughly admitted in the theoretical literature

[64] According to Bowker and Pearson, "language for general purposes [is] the language used by ordinary people in everyday situations. In contrast, a special purpose corpus is one that focuses on a particular aspect of a language. It could be restricted to the LSP of a particular subject field, to a specific text type, to a particular language variety or to the language used by members of a certain demographic group (e.g. teenagers)." (2002, p. 12).

on corpus linguistics; therefore, it should not be seen as the holy grail by corpus-builders, unless it is at the core of their purpose.

For some authors, the idea of sampling or representativeness is a utopia, when the purpose of the corpus is the study of general language usage. As Teubert and Cermáková (2004) argue, it is impossible to have access, and gather all verbal and written texts of a given community of speakers.

Similarly to the previous two authors, Leech (2011) questions the feasibility of this criterion. This author considers this topic as a hard-to-attain criterion

> which tends to suggest an all-or-nothing quality. […T]he latter is something we are optimistically looking for, but may never exactly find. In this respect it is like truth. Very rarely can complete representativeness, like complete truth, be attained. (Leech, 2011, p. 159)

Sardinha (2011), in turn, points out the unawareness of the size of the population as a problematic issue given the possibility of erroneous outcomes from large corpora analysis, therefore, the issue of making generalisations from large samples of language should be handled with caution. For this author, the critical aspect for building a corpus is to include a wide range of text genres in order to represent the population if not totally, at least fairly.

The criterion of *variety of text genres* is corroborated by Laviosa, who also points at this variety as the means of achieving balance: “Representativeness depends on two factors: balance and sampling. Balance is the extent to which a corpus includes the full range of text types that are considered to represent the population.” (Laviosa, 2011, p. 136).

As pointed out above, representativeness is widely debated, and generally seen as the Achilles’ heel of corpora building in Corpus Linguistics (see Leech, 2011). Nonetheless, some corpora may achieve a reasonable degree of this criterion, as stated by Johansson (2011):

> It is particularly problematic to compile representative general-language corpora, where it is virtually impossible to define the population from which a sample is drawn. [However], [t]he more limited the aim, the greater the chance of compiling a well-defined corpus and achieving a reasonable degree of representativeness. (Johansson, 2011, p. 118)

Johansson's words may lead us to conclude that the goal of a given study is what underlies the representativeness criterion for a corpus to be representative of a well-delimited discourse community. This assumption is reinforced by Viana's (2011) view: "the more specialized a corpus is, the easier it is to gather a relevant sample of the language to be studied." (Viana, 2011, p. 232).

As outlined above, when the subject is *special corpora*, some authors point at different measures of criteria for their compilation, in particular regarding *representativeness*. According to Tognini-Bonelli:

> There are […] collections of texts which do not provide [lexico-grammatical] kind of evidence, but which are still referred to as types of corpora. The selection method, or the pool of texts from which the selection is made, is not designed to be representative of a language or variety. Many of these are important collections, and to mark both their importance and their difference from 'ordinary language' corpora they are called special corpora. (Tognini-Bonelli, 2010, p. 22)

These last words corroborate Costa's (2001) perspective concerning the authority of specialised texts to build specialised corpora.

Since our aim is a specialised corpus, we will focus on the criterion *balance* instead of representativeness given the purposes of our study. This option is based on Costa's (2001) perspective : when in presence of a corpus built uniquely with specialised texts, the criterion of representativeness is complemented, "non au sens statistique, mais au sens de l'acceptation du texte en tant que reproduction scientifiquement reconnue par les membres qui composent la communauté scientifique ou professionnelle, dans laquelle et par laquelle le texte a été produit" (Costa & Silva, 2008, p. 7).

Thus, given the recognised authority conveyed by specialised texts, for they efficiently represent the experts' socio-discursive context – production and reception of the text – and coherently mirror the knowledge of the domain they belong to, the criterion of representativeness is considered to be achieved in a specialised corpus.

2.9.5. Balance

Contrary to representativeness, the notion of *balance* is stated as a crucial characteristic, not only in the corpus linguistics literature (Atkins, Clear, & Ostler, 1992; Laviosa, 2011; Gries, 2011), but also in the literature regarding terminological work (Pearson, 1998; Meyer, 2001; Bowker & Pearson, 2002; L'Homme, 2004), among others.

As advanced by Meyer (2001),

> [i]t is well-known in the terminology literature that technical texts correspond to a variety of communicative situations: experts writing for other experts, experts communicating with students of the field, experts or semi-experts writing for the laypublic. [...] Like lexicographers, terminographers try to build "balanced" corpora (Cf. Meyer and Mackintosh 1996), and one way to achieve balance is to ensure that the corpus texts represent a range of communicative situations. (Meyer, 2001, p. 318)

Balance is described by Atkins, Clear, and Ostler (1992) as a *sine que non* criterion for corpus analysis work. According to these authors, a balanced corpus

> is meant (apparently) a corpus so finely tuned that it offers a manageably small scale model of the linguistic material which the corpus builders wish to study. At present corpus 'balance' relies heavily on intuition, although work on text typology is highly relevant. (Atkins, Clear, & Ostler, 1992, p. 14)

Moreover, a predefined text typology is fundamental for corpus compilation. Text is language material and it is the quality of text variety that will qualify the corpus as balanced. Balance is thus considered to be an outcome attested through the corpus analysis, instead of a predefined criterion:

> Controlling the 'balance' of a corpus is something which may be undertaken only after the corpus […] has been built; it depends on feedback from the corpus users, who as they study the data will come to appreciate the strengths of the corpus and be aware of

its specific weaknesses. […] Knowing that your corpus is unbalanced is what counts. (Atkins, Clear, & Ostler, 1992, p. 14)

Hence, an attempt of balance is suggested to be set *a priori.* Corpus builders must predefine a text typology for their corpus design, such as genres of texts, communication settings, and other variables, by means of "classifying the texts which they have chosen in order to facilitate the retrieval of information from the corpus" (Pearson, 1998, p. 52).

Costa and Silva (2008) also make reference to the pertinence of the classification of texts for their terminological work; although, highlighting the text structure as a requirement for the classification of specialised texts, in order to mirror a typology of texts pertaining to a specific domain: "les textes doivent maintenir entre eux des relations de ressemblance au niveau des macro et des microstructures à travers l'identification de régularités propres à un ensemble de textes, par opposition aux régularités d'autres ensembles de textes." (2008, p. 6). According to these authors, a typology of texts is an outcome of the organisation of the collected texts according to their common characteristics; a feature from which a classification is attainable. Such classification allows a systematic distribution of texts in groups or types, to which either a label or a generic name is assigned. This ingathering, which is always artificial and dependent on the researcher's point of view given her/his project goals, may be from the linguistic or the extralinguistic order.

The criterion of text structure regularity is also referred to as an internal criterion as opposed to external criteria, further discussed in the next section.

2.9.6. Internal and external criteria

The types of texts that are assembled to constitute a corpus are selected according to two main criteria: internal and external criteria. According to Pearson (1998), "the emphasis tends, in general, to be on external criteria, both for the classification of texts and for the design of corpora." (p. 53).

However, both internal and external criteria should be considered for the task of text capture, as argued by Atkins:

> A corpus selected entirely on internal criteria would yield no information about the relation between language and its context of situation. A corpus selected entirely on external criteria would be liable to miss significant variation among texts since its categories are not motivated by textual (but by contextual) factors. (Atkins, Clear, & Ostler, 1992, p. 8)

External criteria are those which are essentially non-linguistic and concern text categories, e.g., text genres, participants, communicative function, occasion and social setting. On the other hand, internal criteria concern the lexical and grammatical features of the text (i.e., according to its linguistic characteristics such as the author's vocabulary choices, diction and syntax, from which the text is classified as formal or informal) (see Pearson, 1998*;* Laviosa, 2011).

2.9.6.1. The broad external criteria

According to Pearson (1998), the broad categories of external criteria include (1) genre, (2) mode and (3) origin:

The *genre*[65] category allows for distinctions to be made between different types of written publications, such as books, subdivided into fiction and non-fiction, and so forth. Common genre categories are press, religion, fiction, private letters and academic (see Baker, Hardie, & McEnery, 2006).

The *mode* describes in what form a text was originally produced: either in its original written form, or if is a transcription of a spoken text.

The *origin* indicates who has been involved in the production of a text, i.e., the author, editor, publisher, and so forth.

There is another criterion pointed out by Pearson: the *intended outcome* of the text, which is the purpose for which a text was written and includes the following

[65] The issue of genres will not be discussed in our study. However, Flowerdew interestingly points out that: "Genres operate at the level of discourse structure, which is determined by the communicative purposes of the text and the sociocultural context." (Flowerdew, 2001, p. 25). This assumption corroborates Pearson's perspective of communicative settings.

categories: information, discussion, recommendation and recreation (see Pearson, 1998).

In our terminological work, we will mainly focus on two of those external criteria for corpus design, namely *origin* and *intended outcome*, in particular the sub-types *author, information* and *recommendation* since we are aiming at texts written by experts of the industry of cork.

2.9.6.2. The broad internal criteria

The broad categories of internal criteria are based on (Pearson, 1998):

(1) topic and (2) style.

Topic is a controversial matter as stated by Pearson. However, this author points at some general assumptions of how this matter can be identified:

- by looking at what a particular text is, on the basis, for instance, of its title or its table of contents, and classifying the text accordingly;
- by examining the lexical structure of the text and identifying keywords used frequently.

Style is another controversial matter for Pearson, given the lack of consensus within the corpus community for the categories of *formal, informal or colloquial* to classify text style. However, Pearson argues that there is a tendency to assign a style category based on genre and text purpose. As this author exemplifies, a report is more likely to be classified as formal, while a discussion may be classified as informal or formal.

The list of external and internal criteria stated above is not exhaustive, nor is meant to be a model for corpora compilation.

Based on what has been said so far, we can conclude that what corpus builders must keep in mind as the first assignment in corpus building is to focus on the purpose of the corpus itself to efficiently achieve the goals of the corpus analysis they have envisioned. Motivated by the purpose of the corpus, internal and external criteria for

text classification should be well predefined by the corpus-builder in order to underpin the desired outcomes of the corpus, e.g., if the purpose of the corpus is to analyse certain linguistic regularities such as morpho-syntactical patterns suggesting special knowledge information, one of the external criteria to be thought of is the *technicality* (Atkins, Clear, & Ostler, 1992; Costa R. , 2001) of the texts. *Technicality* classifies a text according to the degree of the author's knowledge and the targeted audience's knowledge (e.g., a research paper is an expert-expert setting of communication). While internal criteria, which are not as obvious as external criteria, should be put in place through the observation of the text itself, where the author denotes expertise of the matter under discussion – the *topic* – in a well-structured discourse by means of the coherence[66] and cohesion[67] of his lexical-grammatical choices throughout the whole text.

The lexical-grammatical choices of the experts of a given domain and the specialised context of production vs. reception, also known as *communicative setting* (Pearson, 1998), are considered fundamental to observe specialised discourse; therefore, pointed out as crucial criteria for terminological studies (see Atkins, Clear, & Ostler, 1992; Pearson, 1998; Costa R. , 2001). Hence, and inspired in these authors, the criterion *communicative setting* is at the core of the design of our domain-specific corpus, as discussed in Section 3.1.2. (p. 79).

[66] is denoted by "les faits de continuité et de progression sémantiques et de référentielles produits dans un texte, ou plus largement dans le discours, par un dispositif spécifiquement linguistique. Le terme vise donc l'ensemble des moyens mis au service de la liaison intraphrastique et interphrastique […] et qui permettent à un énoncé d'apparaitre comme une séquence textuelle ou discussive formellement unifiée." (Neveu, 2015, p. 85).

[67] is conveyed by the "propriétés pragmatiques qui assurent à une séquence textuelle ou discursive son interprétabilité, notamment par des données informationnelles (portant sur des actions ou situations) susceptibles d'être congruentes avec le monde de celui qui évalue ces données. On fait généralement entre dans ces propriétés des connaissances culturelles, des valeurs morales ou idéologiques, des lieux communs, etc." (Neveu, 2015, p. 85).

Corpus of analysis

3. TermCork: A corpus-based research to perceive domain-specific concepts

Concepts[68] along with terms and definitions are at the core of terminological research. According to ISO 704 (2009), the terminological practice is perceived as being the analysis and processing of concepts and terms from a given scientific, technical and/or professional domain. Terminologists (linguists), however, do not master the concepts of a given special field, and neither its terminology unless they are experts in that same special field under analysis. In this context, terminologists are thus non-experts by definition. Which leads one to formulate an obvious question: How can a non-expert terminologist have accurate access to the knowledge – to the concepts – of a special field by means of corpus analysis?

The answer to this question will be provided by our research for which we decided to build a domain-specific corpus, i.e., a corpus containing texts from a single domain and for a specific terminological purpose: to analyse the discourse of experts in order to have access to their conceptualisations[69].

[68] "représentation mentale qui retient les caractéristiques communes à un ensemble d'objets. Les objets du monde réel sont tous différents mais il est raisonnable de penser que la représentation que nous nous en faisons retient l'essentiel de leurs caractéristiques, ce qui nous permet d'en reconnaitre de nouveaux. En terminologie classique (à optique conceptuelle), cette représentation mentale est donnée comme posée (c'est à dire qu'on ne cherche pas à en expliquer la nature) et on considère qu'elle précède la forme linguistique comme telle." (L'Homme, 2004, p. 25).

[69] According to Pottier, "La conceptualisation [est une opération] préverbale [et] permet de choisir un type d'évènement […] et de choisir également les aspects du référent qui seront retenues. […] Le résultat est le *message,* toujours unique […] puisqu'il n'est jamais totalement reproductible et que sa composante implicite n'est pas totalement identifiable." (1992, pp. 14-15).

Our corpus-based[70] research is not much different from the abovementioned terminological studies. Pearson states that lexicographers and terminologists share the same corpus-approach:

> The corpus-driven approach is likely to be used by lexicographers, terminographers and computational linguists when they are seeking to discover new facts about a language. The Cobuild dictionary is a product of the corpus-driven approach to lexicography. The meanings of words are identified by means of an analysis of their usage in text. Terminographers may use the corpus-driven approach to identify potential terms in a corpus. (Pearson, 1998, p. 50)

In our terminological work, one of the goals is *to identify the meaning of words by means of analysing their usage in text*; yet it is not merely a word; on the contrary, it is a term and first and foremost the concept that is designated by that term. In other words, as terminologists, we search for *designations*[71] that acquire mono-referential meaning when used in a special context of communication, i.e., terms pointing at concepts.

Hence, the focus of our terminological project is not the study of general language in its every-day-usage, but the terminological choices made by the interlocutors in a special context of communication: the *discourse* of experts as previously stated. Some authors refer to this *discourse* as a Language for special

[70] According to Baker, et al., the author "Tognini-Bonelli (2001) makes a useful distinction between corpus-based and corpus-driven investigations. The former uses a corpus as a source of examples to check researcher intuition or to examine the frequency and/or plausibility of the language contained within a smaller data set. The researcher does not question pre-existing traditional descriptive units and categories. A corpus-driven analysis is a more inductive process: the corpus itself is the data and the patterns in it are noted as a way of expressing regularities (and exceptions) in language. A corpus-driven analysis tends to only use minimal theoretical presuppositions about grammatical structure." (Baker, Hardie & McEnery 2006, p. 49).

[71] In the sense of ISO: a "representation of a concept by a sign which denotes it in a domain or subject. Note 1: A designation can be linguistic or non-linguistic. It can consist of various types of characters, but also punctuation marks such as hyphens and parentheses, governed by domain-, subject-, or language-specific conventions. Note 2: A designation may be a term including appellations, a proper name, or a symbol." (ISO/FDIS 1087, 2019 (E), p. 7).

purpose[72] (LSP), and others as specialised discourse. The dichotomy language vs. discourse is widely debated among scholars, but it will not be a matter of discussion in this study. Our terminological choice over this dichotomy follows the notion of discourse, an outcome of the linguistic choices observed in a community of speakers, as advanced by Teubert and Cermáková:

> [corpus linguistics] can tell us more about the meaning of words than standard or Chomskyan linguistics. It extracts from the discourse all that we can find about meaning. Natural human language is unique in that respect. It is the discourse community that negotiates how words should be used and what they mean. The result of these negotiations is not always agreement. [...] We only have to look at the recent discourse to find numerous citations in which people are keen to tell us what they think [a certain word/expression] are [...], and, consequently, what the phrase [in which they occur] means. (2004, p. 105)

In a sort of analogy, if the discourse community negotiates the meaning of words belonging to the general language, we can assume that the *discourse community*[73] of a special field that negotiates the meaning of terms – the experts – is a community of specialised discourse, as pointed by Costa (2001).

To approach this specialised discourse, our corpus compilation requires, as an essential criterion, the inclusion of texts produced by experts where discourse is governed by the pragmatic constrains of the context: the higher the expertise of the interlocutors, the more specialised the discourse. As mentioned before, the

[72] Other authors clearly differentiate special from specialised: “In our understanding, a term is a “special” sign; “special”, i.e. as we speak of a “special domain” or, in French, of "gens spéciaux" (‘special people’), i.e. people at work in their own speciality. Lastly, we define “Language for special purpose” as a bundle of units – terms, words, expressions – and combination rules, which comprises a whole language used in a domain of knowledge. There are many ways to operate with that simple distinction.” (Depecker, 2015, pp. 38-39).

[73] Teubert and Cermáková define discourse as an outcome of the totality of verbal productions: “What exactly is the discourse? A language, a discourse, consists of the totality of verbal interactions that have taken place and are taking place in the community where this language is spoken. This community we call the discourse community.” (Teubert & Cermáková, 2004, p. 114).

communicative setting[74] is a core-criterion of our search for and capture of texts, for a domain-specific corpus compilation, although not exclusively, as discussed in the next section (3.1). The underlying expectation of this criterion is to create a text collection that coherently represents the social-discourse community of the professional, technical, and scientific practice within the special field of cork.

3.1. Domain-specific corpus: purpose and design

To attain our terminological goals, we decided to build a domain-specific corpus, i.e., a corpus comprised of texts produced in a specialised context of communication, where the discourse of a community of experts from a field of interest is reflected. The overall purpose of the creation of this corpus is to analyse the discourse of experts in order to extract information that represents the experts' conceptualisations beyond their verbal expression.

Texts are undoubtedly vehicles of knowledge transfer. Therefore, one of our tasks is to analyse texts in order to subtract the concepts' linguistically expressed characteristics, which will allow us to grasp conceptual relations that are specific to the domain. This will permit us to propose a preliminary conceptual organisation of the subject field to be discussed later on with experts.

3.1.1. Corpus criteria design: text type, format and publication date

As the first criterion within the task of building a domain-specific corpus that mirrors the industry of cork, we decided to collect texts written by experts in order to effectively represent this domain through their topic and communicative purpose. The topic is restrained to the domain of cork, with major emphasis on cork stoppers, and the communicative purpose of the text is preferably the information or normalisation type, but not exclusively.

[74] Following the perspective of Pearson: 1) expert-expert; 2) expert-initiates; 3) relative expert-initiate; 4) teacher-pupil (1998, p. 38).

Secondly, we have decided to capture texts from the internet since their electronic format is machine-readable. However, we found at the Portuguese National Library (BNP) some texts, in hard copy, that we considered relevant to be included in our corpora collection given their authoritative authors, thus a few hard copies of texts had to be digitalised and subsequently ran through optical character recognition (OCR) tools, in order to acquire the desirable e-format for machine readability.

Thirdly, the date of publication of the texts should not exceed 10 years, so we could observe the experts' terminological choices synchronically. The time frame of ten years has to do with the rapid evolution of technical domains.

We must note that the corpus compilation started in the year 2013[75]. By that time, the time frame of ten years dictated that the text's publication dates should belong to the period of 2003-2013. However, we found ourselves forced to widen the time frame because some concepts of the domain were not defined in more recent texts; thus, a text published in 2001[76] was included in our text collection. These considerations about older texts were also pointed out by Bowker and Pearson (2002):

> older texts can also be valuable: experts usually provide lots of definitions and explanations when a new concept is developed, or a new term is introduced, but these explanations become less frequent as this information becomes part of the experts' general knowledge. (Bowker & Pearson, 2002, p. 52)

Motivated by this need of widening the time frame, we decided to continue our text capture to feed our corpus while the project was ongoing, albeit without discarding any older text. With this last decision, we have characterised our corpus as a dynamic one – a non-finite corpus – given the continuous addition of new texts. We must stress though that the publication date of the new additions should always respect the time frame of 10 years, except in case of fundamental reasons, as mentioned above.

---

[75] See (Ramos, 2015).

[76] An industrial technical guide for the industry of cork issued from the National Institute of Engineering, Technology and Innovation (INETI), in 2001: "Guia Técnico Sectorial - Indústria da Cortiça".

3.1.2. The communicative setting

The scarce availability of and accessibility to texts produced by the discursive community of a special field increases the difficulty to achieve the desired balance of a corpus. This assumption led us to agree the more genres of texts, the better, as long as the predefined criterion of communicative settings of Pearson (1998) was complied with. In our study, the predefinition of the communicative setting is a variable within the external criteria for text search and capture.

Thus, from the specific criteria we predefined for the collection of texts to be included in the corpus, we particularly focused on the communicative settings of production/reception, where authorship is of utmost importance for the reliability of the information contained in the texts and the intended outcome of the linguistic analysis. The texts were compiled according to the following criteria:

1. texts produced by and for the scientific community of the domain of cork;

2. texts produced by experts for quasi-experts;

3. texts produced for non-experts.

The rationale behind the inclusion of the third group in the corpus is the fact that these texts are rich in definitional contexts[77] and/or contexts[78] that describe concepts given the different degrees of knowledge of producers and recipients.

Following the three criteria mentioned above, we obtained a balanced corpus that covers the different levels of specialised discourse.

In sum, the communicative setting of the production of the texts was the most significant criterion for the compilation of the corpus to support our terminological purposes. An important aspect is that the linguistic analysis aims at observing texts produced by experts for semi-experts or quasi-experts that are commonly technical-

[77] By definitional contexts, we mean contexts that are rich in knowledge information permitting the elaboration of definitions (Ramos, Costa, & Roche, 2019).

[78] A fragment of text that helps to explain the meaning of a linguistic expression.

explanatory, as well as normative texts, and texts produced for the economic and financial areas (the latter were produced by experts of the domain for experts of governmental institutions). The underlying reason for this option is that these texts contain glossaries and definitions produced by experts; thus, validation of the extracted terms[79] is provided *a priori*. The remaining corpora are used as reference corpora.

The abovementioned internal and external criteria are systematised below in Table 6.

Table 6: Internal and external criteria of the cork corpus

Criteria	***Purpose/description***
Degree of specialisation	Produced by and for experts
Source validation	Entities recognised as an authority
Type	**Technical-explanatory; normative**
Content adequacy	On Cork/Cork stopper
Synchronism (≤ 10 years)	Given the fast evolution of technology

Table 6 represents the internal and external criteria we have predefined for text capture in order to constitute a collection of texts that we can classify as specialised corpora, thus constituting a domain-specific corpus.

In addition to these criteria, the language of the texts was also determined.

3.1.3. The language eligibility criteria

We have captured texts in three languages, in order to compile a multilingual corpus. These languages are:

European Portuguese (PT), the language of the main producer of cork in the world;

French (FR), the language of the greatest client of Portuguese cork;

[79] The analysis of these glossaries and definitions have been at the core of our terminological work since 2013. Thus, a considerable amount of terms and definitions of the domain has already been compiled.

English (EN), the language commonly used as *lingua franca* for international business.

One of the interesting aspects during the task of text capture for their inclusion in the multilingual corpus was finding several identical texts, issued by the same institution[80] – whose author is a recognised authority in the domain – and translated in these three languages. With these texts, we were able to create a *parallel corpus*[81] and observe how a given concept is designated in the three languages; therefore, equivalents of some terms are identified. However, the expert will have to validate these equivalents, in a later moment[82], since translation is not a task of experts, but of language specialists.

The purpose of these three collections of texts is to observe how concepts are designated in the three different languages. By means of corpus analysis, terms and corresponding equivalents may be seen in context, a feature that will allow us to produce terminological resources, like for instance, glossaries, or (re)write terminological definitions in which equivalent terms play a fundamental role. In our opinion, such terminological resources are an added value for the improvement of international communication purposes, or for specialised translation, or in our case, for the creation of a multilingual special field dictionary within the scope of cork.

3.1.4. Composition of the text collection, written in Portuguese

The pt corpus is comprised of 98 texts written in European Portuguese. These 98 texts were produced by experts belonging to different organisations coming from different areas — scientific, industrial, techno-professional, certifying, regulating, and commercial — and are available online, except for a few chapters of two books that we

[80] APCOR© - Associação Portuguesa da Cortiça [Portuguese Association of Cork].

[81] This kind of corpus "contain texts and their translations into one or more languages. A bilingual parallel corpus contains texts and their translations into one language, and a multilingual parallel corpus contains texts and their translations into two or more languages." (Bowker & Pearson, 2002, p. 92).

[82] Expert's validation is not contemplated in this project. Yet, future work will require "mediation strategies between terminologists and experts" (Costa, Silva, Barros, & Lucas Soares, 2012).

found in the repository of the Portuguese National Library (BNP), in hard copy (as stated in Section 3.1, p. 77).

Table 7: Corpora collection – 98 texts produced following 3 major criteria: expert-expert; expert-quasi-expert; expert-non-experts.

corpus n=98	**technical**	**legal**	**scientific**	**economics**	**marketing**	**arts & history**
Specialised periodicals	3			1		
Books	3		1			4
Instruction manual	4					
Industrial guide	1					
Standards	9					
Decree-law		6				
Theses			13			
Academic articles			15			
Reports				8		
Studies				7		
Brochures					6	
Newsletters	7				10	
Total	**27**	**6**	**29**	**16**	**16**	**4**

We have recorded in Table 7 the types of texts we have collected according to the predefined criteria we have mentioned above.

On the vertical axis of Table 7, we have organised texts based on an external criterion – the purpose of the text according to its intended audience – while on the horizontal axis, texts were organised according to external vs. internal criteria – the author's linguistic choices according to the communicative setting vs. the intended outcome of the text (i.e., explanatory; regulatory; scientific; informative).

The intended outcome identified in the 98 texts is listed below, each corresponding to a distinct communicative setting of production/reception, as noted after the arrow:

- explanatory/normative (technical) → expert – quasi-expert / professionals
- regulatory/prescriptive (legal) → semi-expert – expert
- scientific/educational/disclosure (scientific) → expert – expert
- economics – informative (economics) → expert – semi-expert
- promotional (marketing) → semi-expert – non-expert
- narrative - informative (arts & history) → semi-expert – non-expert

The communicative setting semi-expert-non-expert is not under our terminological focus, for texts produced in this communicative setting are not rich in knowledge patterns or definitions. This type of information is usually conveyed by the experts' discourse. However, we have considered these texts useful to be used as reference *corpora*[83], i.e., a wider text typology that serves to clarify the doubts of the linguist – a non-expert by definition – such as, for instance, to observe the different morphological structures of a given term used in non-expert settings of communication vs. expert-semi-expert; and/or how the designated concept is described or explained in non-expert communicative settings.

Similarly to the previous communicative setting, the expert-expert and semi-expert-expert settings are also not under our primary terminological focus. Experts master the concepts of their domain of expertise. Consequently, concepts are unlikely defined in these two settings of text production/reception, unless when a new concept arises in the domain. Since the goal of our corpus analysis is mainly the capture of

[83] In the sense of "when using frequency-based techniques to analyse a text or set of texts, it is necessary to have something with which to compare them. This is necessary, for instance, if we wish to establish that some word or form is more common in a particular text than is normally expected. The basis for the comparison is often a larger set of texts drawn from a wider range of genres and/or sources. This larger dataset is often called a reference corpus." (Baker, Hardie, & McEnery, 2006, pp. 136-137).

concept definitions, these texts will also be used as reference corpora, for in spite of the lack of definitions, these texts abound in terminology.

3.2. The corpus of analysis

The corpus of analysis – the set of texts over which our terminological analysis was performed – is comprised of the collection of texts that we classified as pertaining to the quasi-expert communicative setting. Such texts are produced in two communicative settings, namely expert – semi-expert and expert – quasi-experts / professionals, as shown below in the graph (Figure 8). As mentioned before, the reason for this option is tied with the definitional (con)texts produced by experts, e.g., glossaries and definitions. These definitional (con)texts are commonly found in these two settings of communication, for the more significant the knowledge gap between the author-expert and his/her audience, the more definitions and contextual definitions are produced.

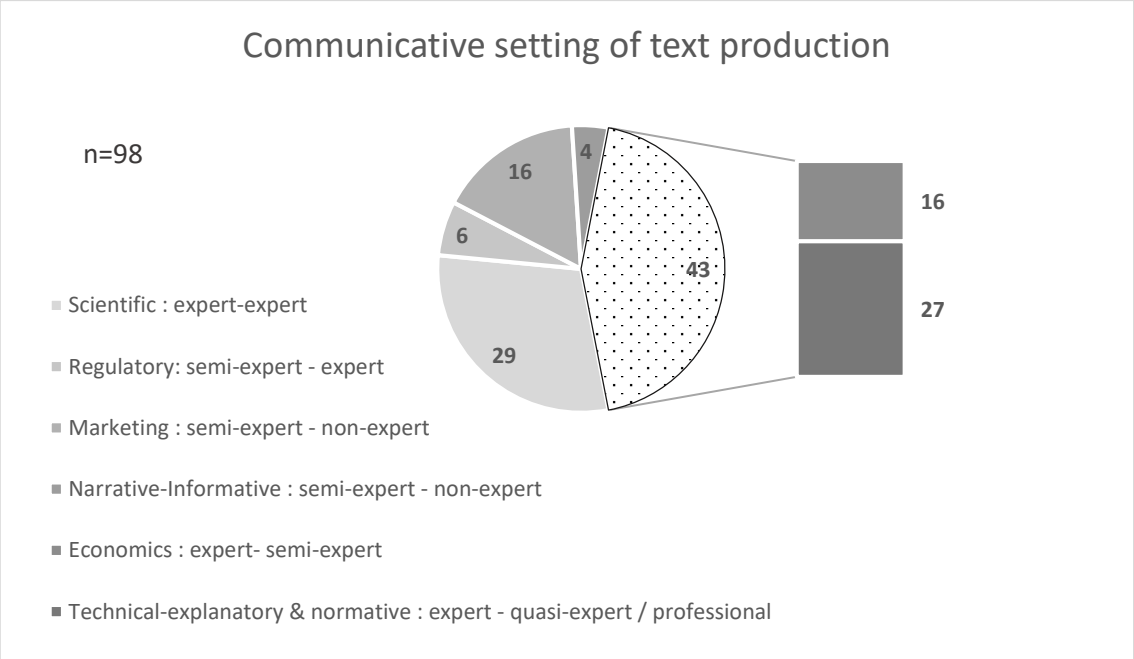


Figure 8: Corpus of analysis based on the communicative setting of expert – semi-expert (Economics) and expert-quasi-experts / professionals (Technical-explanatory).

The terminological analysis focused thus on 43 texts produced in the two mentioned communicative settings, as shown in the graph above (Figure 8) while the remaining 55 texts were used as reference corpora. As we can observe in the graph, the quasi-expert communicative setting unfolds in 16 texts produced in the Economics setting (expert – semi-expert) and 27 in the Technical-explanatory & normative (expert – quasi-expert/professionals).

3.2.1. Composition of the multilingual text collection

Regarding the two other languages we decided to work with, namely French (fr) and English (en), the fr corpus and the en corpus are two collections of texts composed of the following types:

Table 8: Types of the FR and EN corpora

TEXTES	No.	langue	TEXTS	No.	language
BROCHURES	4	fr	BROCHURES	4	en
DÉCRET-LOI	1	fr	DECREE-LAW	0	en
TECHNIQUES-EXPLIC.	6	fr	TECHNICAL-EXPLAN.	2	en
ÉTUDES	0	fr	STUDIES	6	en
LIVRES	5	fr	BOOKS	3	en
NORMES	13	fr	STANDARDS	10	en
ARTICLES ACADEMIQUES	8	fr	ACADEMIC ARTICLES	21	en
BULLETINS	22	fr	NEWSLETTERS	11	en
RAPPORTS	0	fr	REPORTS	1	en
THÈSES	1	fr	THESES	3	en
TOTAL	**60**			**61**	

As seen in Table 8 above, the number of texts composing the fr and en corpora are not as large as the pt corpora. The predefined criteria for their capture were identical to the design of the pt corpus, in which the communicative setting of production/reception is the major criterion.

However, the purpose of these two corpora is not identical to the pt corpus.

The fr and en corpora were not built to be analysed in the same way as we did with the pt. While the latter was explored to analyse the experts' terminological choices in order to grasp their conceptualisations, the other two were used as comparable

corpora[84] and/or parallel corpora for the search of equivalent terms, in fr and/or en, for a given term in pt.

3.2.1.1. Multimodal corpora

In addition to the multilingual text collection, we also found a collection of images and videos regarding different activities within the field of cork, namely the production of cork in forests and/or the industrial products made from cork and their corresponding line of manufacture, just to state a few. These images and videos are also produced in the three languages we are working with and are available on the internet.

Our interest for these multimedia files has to do with the possibility of being associated to the definition of a given concept, in a terminological resource, in which the text of the definition is complemented with an image or a video. This interest was inspired on the words of Rey-Debove:

> [...] la définition qui remplit sa fonction abstraite d'identification est insuffisante pour évoquer l'objet. D'abord, parce que les traits pertinents qu'elle propose sont différentiels plus que positifs (distinguer l'âne du mulet, [...]), ensuite, parce que les traits liés à l'aspect visuel sont parfois secondaires et que néanmoins ce sont ces traits-là qui nous aident à identifier l'objet. [...] l'image fonctionne plus comme un signal que comme un ensemble de traits. (1998, p. 272)

In a terminological resource, the central role is played by the definition of the concept, either through a written text or formal schema, or another form of representing a definition depending on the domain under focus, such as, for instance, a mathematical formula, for "[...] l'un des avantages de la définition sur l'image, [...] est sa pérennité due à sa plus grande abstraction" (Rey-Debove, 1998, p. 272). However, we believe that multimedia files still have added value for non-experts users of the terminological resource, so they can effectively infer the concept, because in spite of

[84] In the sense of Bowker and Pearson: "corpora consist of sets of texts in different languages that are not translations of each other. We use the word 'comparable' to indicate that the texts in the different languages have been selected because they have some characteristics or features in common; [...] The shared features will frequently include subject matter or topic and may also include features such as text type, period in which the texts were written, degree of technicality, etc." (Bowker & Pearson, 2002, p. 93).

"[l]es objets fabriqués changent d'aspect et rarement de fonction; un dessin d'automobile vieillit entre deux salons." (*ibid.*). According to this author, industrial domains benefit from the usage of images given the rapid evolution of the technology; thus, from this assumption, we can assume that a terminological tool focusing on the industry of cork – an industry and/or products in constant evolution – will benefit from the use of multimedia files along with the textual definitions.

We can conclude that we have built a *specialised corpus*, in the sense of Costa (2001), since this corpus is composed of texts produced by a professional class, as well as compiled in a significant number, so that the specialised statements are fairly representative of that community of expertise. The novelty here is its multilingual and multimodal aspect, i.e., it is a corpus built with a collection of texts produced in three languages and in different semiotic mediums, i.e., in written and image (fixed or in motion) forms.

The goal of such multimodal corpus is to create a terminological knowledge database (TKB) to feed a multisemiotic e-dictionary given its multimodal resources.

3.3. Corpus management

For corpus management, we have resorted to the Sketch Engine[85] corpus software package.

We used Sketch Engine to compile, annotate, and query the corpus employing a Corpus Query Language format, where REGEX[86] are used. Furthermore, this tool has an

[85] https://www.sketchengine.eu/

[86] A regular expression "is a compact way of describing complex patterns in texts. You can use them to search for patterns and, once found, to modify the patterns in complex ways. They can also be used to launch programmatic actions that depend on pattern.": http://gnosis.cx/publish/programming/regular_expressions.html [Accessed 06-07-2020].

incorporated tagger for Portuguese called FreeLing, whose tagset[87] was used for the REGEX queries.

Sketch Engine is a tool developed by Lexical Computing[88] – a research company founded by Adam Kilgarriff in 2003. This corpus software allows corpus management and corpus query, where one of the standout features is Word Sketch:

> The word sketch processes the word's collocates and other words in its surroundings. It can be used as a one-page summary of the word's grammatical and collocational behaviour. The results are organized into categories, called grammatical relations, such as words that serve as an object of the verb, words that serve as a subject of the verb, words that modify the word etc. (Lexical Computing, 2020)

We have systematised below, in Figure 9, the most relevant features of Sketch Engine (SKE), a few of which we have used in this study.

[87] "A tagset is a list of part-of-speech tags (POS tags for short), i.e. labels used to indicate the part of speech and sometimes also other grammatical categories (case, tense etc.) of each token in a text corpus": https://www.sketchengine.eu/portuguese-freeling-part-of-speech-tagset/?highlight=freeling.

[88] https://www.lexicalcomputing.com/lexical-computing/

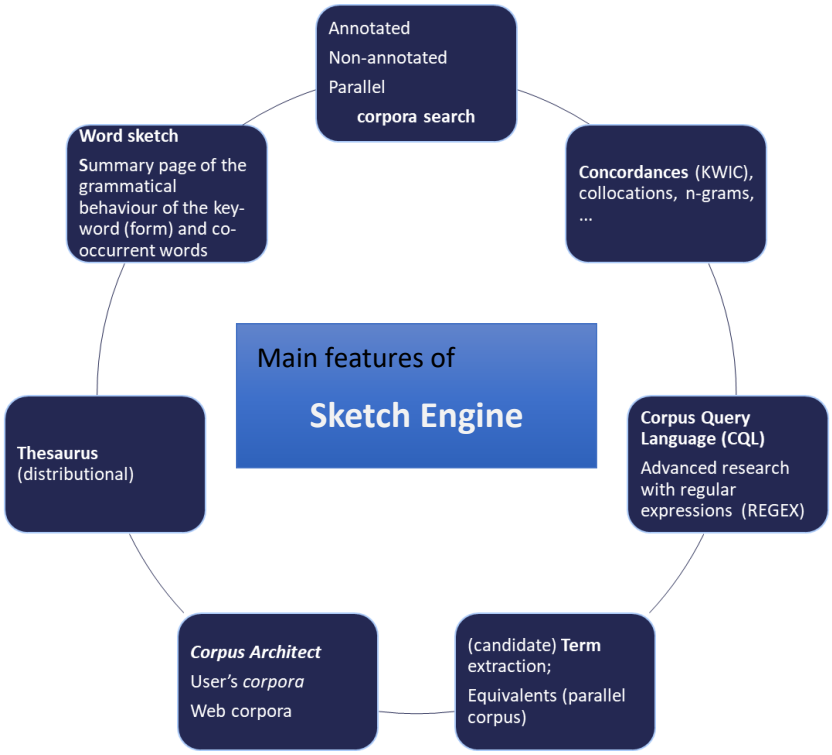


Figure 9: Main features of Sketch Engine

From the above systematised features in Figure 9, we have resorted to the following ones to manage the corpus, in the order enumerated below:

(1) corpus architect

(i) user's corpora: our own texts collected from the internet;

(ii) web corpora (a corpus automatically crawled from the web through WebBootCat[89] – a web crawling[90] feature;

[89] For more details see http://bootcat.dipintra.it/

[90] This feature is related with the "structure of the World Wide Web [which] can be viewed as a directed graph, where everything is present in a hierarchy. When a page is visited, it contains links to other pages. While viewing the Internet as a directed graph, web pages can be considered as nodes and the hyperlinks can be considered as edges. So, we can summarize the search operation as traversing a directed graph.

(2) corpus annotation with the Portuguese FreeLing part-of-speech tagset[91];

(3) word sketch: a summary page of the grammatical behaviour of the keyword (form) and co-occurrent words;

(4) concordances, which are "a list of all examples of the search word or phrase found in a corpus, usually in the format of a KWIC [key word in context] concordance with the search word highlighted in the centre of the screen and some context to the right and to the left" (Lexical Computing, 2020);

(5) corpus query with CQL (Corpus Query Language), which is an advanced search of the corpus resorting to Regular Expressions (REGEX).

Points (3), (4) and (5) were used in an iterative manner, for as soon as a given candidate[92] term or definition is identified in the textual data drawn from the corpus, either through the analysis of KWIC concordances or the tool's answers – as a result of a given CQL – the process of search restarts, and does not necessarily follow the same order.

Given the two types of corpus architecture we have resorted to, namely user's corpora and web crawled, we have a corpus we can call a hybrid corpus. However, the two types are searchable separately, which in our view is a positive aspect since web corpora tend to be noisy concerning unwanted data, also called boilerplates[93] – a typical

Following this linked hierarchical structure, a web crawler can start with a given page and then visit to all those pages whose links are given in that page. For this way of traversing or crawling the graphical net like structure, they are also known as spiders, and because this process is automated, these web crawlers are also known as robots." (Chatterjee & Nath, 2017, p. 6608).

[91] This tagger "is based on the proposals by EAGLES, which intends to enable encode all existing morphological features for most European languages." Available online at (Lexical Computing, 2020).

[92] In the sense of Bowker and Pearson: "When we speak about 'term candidates', we mean words or phrases that appear to be terms." (2002, p. 145).

[93] it is "known to cause problems if included in text corpora. The frequency count of some terms, such as *home, search, print,* is highly increased giving biased information about the language. Also, hits within boilerplate may be annoying when searching in corpora since they often provide no useful evidence about the phenomenon being investigated." (Pomikálek, 2011, p. 19).

outcome from the "robot" of the web crawling feature. Therefore, our web corpus served us as a reference corpus.

Hence, the corpus we have built, which we currently continue managing for size expansion and metadata edition (i.e., labelling each file according to the type of text, source and language), is an open annotated specialised corpus, for texts are (and will continue to be) selected according to criteria for domain-specific corpus design, as mentioned in Section 3.1 (p. 77).

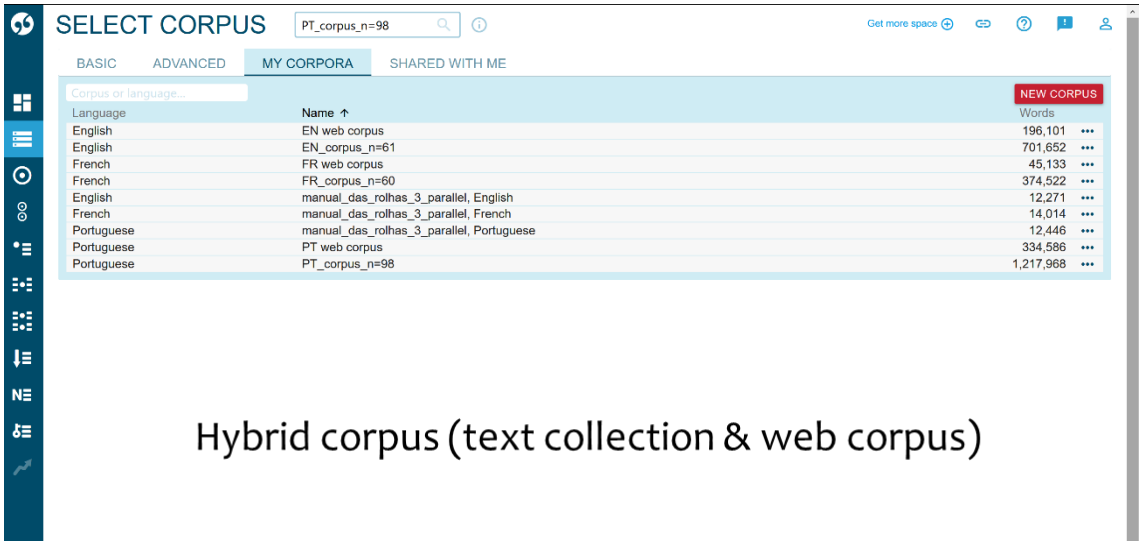


Figure 10: "My corpora" built up via web crawling and user's texts in the Sketch Engine interface

Figure 10 above represents the interface of Sketch Engine, and more specifically our own corpora, which is kept in the tool's cloud. As depicted, the corpus unfolds in several sub-corpora depending on the language and/or the method used for text capture in the internet, i.e., semi-automatically web crawled or manually collected from the internet along with hard copy digitalisations. By semi-automatically web crawled we mean that we parametrised the tool regarding which web pages, and/or sub-pages, the "robot" should explore in order to minimise unwanted data.

Further on, in Figure 10, it is possible to see three parallel corpora, for we had the chance to find a text, originally written in Portuguese, translated into English and French, publicly available on APCOR's page – the Portuguese association for the cork industry.

The process of building the parallel corpus will not be addressed in this study, nor the exploration of both the English and French corpora. As mentioned before, these corpora were useful to search for equivalents; however, for economy of space we will not further develop the topic.

Therefore, the corpus we will address henceforth is the Portuguese corpus, within which the corpus of analysis is nested.

3.4. Corpus processing

Considering the 98 documents of the pt corpus, we have obtained the following quantitative data:

Table 9: Quantitative data of the PT corpus

Frequency [94]	
Tokens[95]	1,712,652
Words	1,217,968
Sentences	48,031

From the observation of the words identified above, we have seen that the most frequent forms that correspond to terms in the domain under analysis are "cortiça" [cork] and "rolha" [stopper], as shown in table (10):

Table 10: The most frequent noun-forms (within the first 300 forms of the list) that correspond to terms in the domain under analysis

Forms (noun)	**English (literal translation)**	**Frequency**	**Percentage per million**
cortiça	***cork***	**16,127**	**9,416.40**
rolha	***stopper***	**5,862**	**3,422.76**

[94] also known as "[…] absolute frequency) refers to the number of occurrences or hits. If a word, phrase, tag etc. has a frequency of 10, it means it was found 10 times or it exists 10 times. It is an absolute figure. It is not calculated using a specific formula.": https://www.sketchengine.eu/my_keywords/frequency/

[95] A *token* is a "single linguistic unit, most often a word, although depending on the encoding system being used, a single word can be split into more than one token, for example he's (he + 's)." (Baker, Hardie, & McEnery, 2006, p. 59).

produto	*product*	4,329	2,527.65
material	*material*	1,667	973.34426
sobreiro	*cork oak*	1,474	860.65354
aglomerado	*agglomerate*	1,228	717.01665
operação	*operation*	1,223	714.0972
rolhas	***stoppers***	**1,221**	**712.9294**
prancha	*plank*	1,017	593.8159
disco	*disc*	777	453.682
granulado	*granular*	705	411.64
preparação	*preparation*	690	402.88
tratamento	*treatment*	640	373.68
pó	*dust*	599	349.74
corpo	*body*	497	290.19
matéria-prima	*raw material*	464	270.92
acabamento	*finishing*	380	221.87
amadia	*amadia*	377	220,12
natural	*natural*	328	191.51

Along with “rolha” [stopper] and “cortiça” [cork], we can see in Table 10 several other terms we have identified within the first 300 forms extracted with SKE – in this case, the tool was parametrised to extract nouns, in a simple list of all the words that fit in this grammatical category criterion. Considering the highest frequency of those two terms and consequently the importance they have in the domain under analysis, we shall look at their behaviour in texts. It must be noted that given our domain-familiarisation, these simple morphologically structured terms – in opposition to polylexical units – were easily identified.

Finally, and as a remark, we can see, on the bottom line of Table 10, the form “natural” [natural] identified by the tool as a noun, instead of an adjective – one of the drawbacks of the POS FreeLing tagger.

In order to identify polylexical terms, we have resorted to another type of word list, in which we parametrised the tool to capture adjectives. The underlying rationale for this grammatical category is tied with the notion of “rolha” [stopper] as a manufactured object; thus, different states of manufacture and/or types are uttered in discourse regarding this object. These different states and types are commonly conveyed in discourse by means of qualities or attributes that take the form of adjectives at the morphosyntactic level given their property of noun modifiers.

The results of the most frequent adjective-forms – within the first 300 forms of the result – can be seen in the next Table 11, and are sorted by lemma – as a default option of the tool:

Table 11: The most frequent adjective-forms (in the first 300) that correspond to terms (or part of polylexical terms) in the domain under analysis

Forms (adjectives)	**English (literal translation)**	**Frequency**	**Percentage per million**
natural	*natural*	1,944	1,135.08
técnico	*technical*	510	297.78
seco	*dried*	434	253.40
cilíndrico	*cylindrical*	207	120.86
lenticular	*lentiform*	161	94.00
suberoso	*subereous*	160	93.42

As systematised above in Table 11, we identified a few adjectives that we hypothesise as being parts of terms, when the latter have a polylexical structure. Once again, we can see the form “natural” [natural] pointed by the tool, but, this time, as the second most frequent form with the grammatical category of adjective, in the entire corpus right after “superior” [superior] – a form we did not include in this list.

Finding adjectives pertaining to the morphological structure of polylexical terms is not an obvious task for non-experts of the domain. Therefore, we decided to use the feature Word sketch, set to the term “rolha” [stopper] and executed only on the corpus of analysis composed of the 43 normative/technical texts. The whole set of results of this feature is depicted below, in Figure 11:

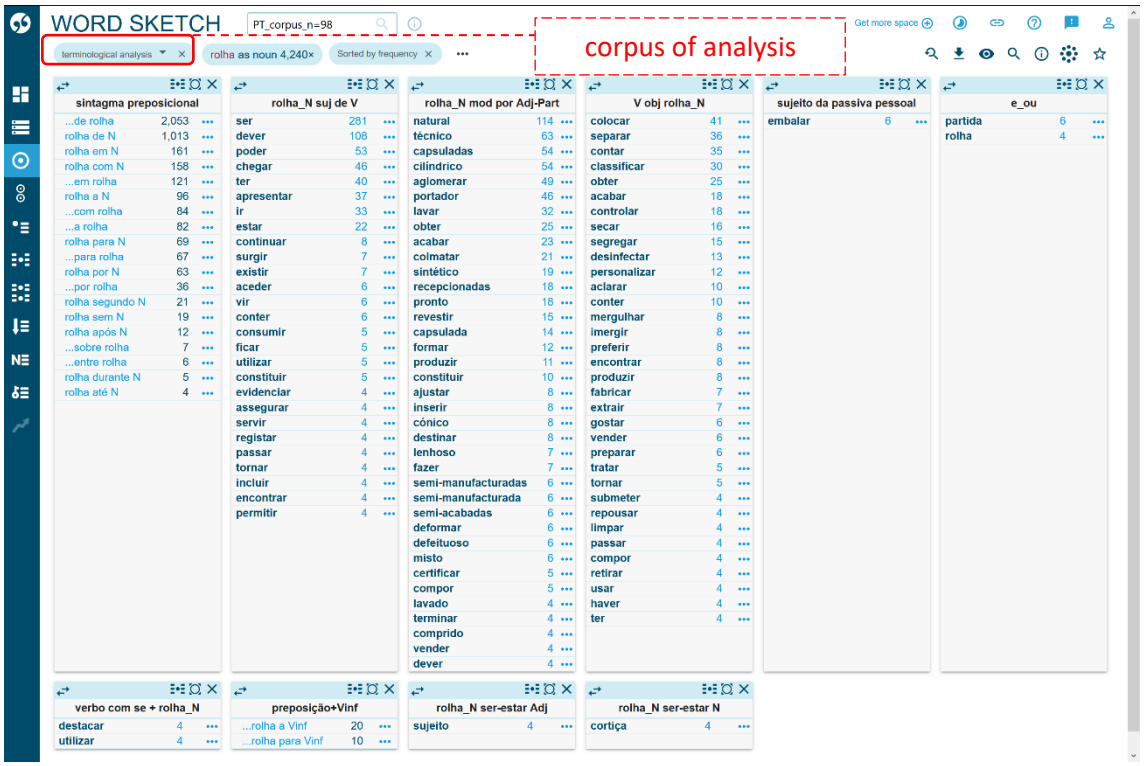


Figure 11: Word sketch for "rolha" [stopper].

To start with, the feature Word sketch gives us a good panoramic view of the keyword's context. This means that we can identify forms that tend to co-occur near and/or in co-text with the KWIC, which provides us with the possibility of observing recurrent morphosyntactic patterns.

From the analysis of the results obtained through the Word sketch for "rolha" [stopper], we could observe the following recurrent information:

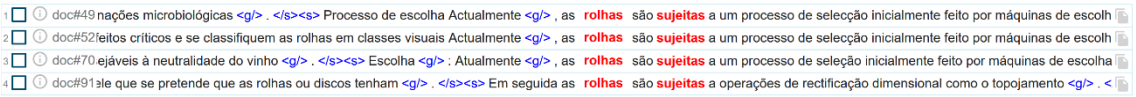


Figure 12: Concordance of rolha_N ser-estar Adj (4 occurrences).

Figure 12 depicts the concordance of "rolha" [stopper] as a noun co-occurring 4 times with the lemma of the verb "ser" or "estar" [to be] along with an adjective, in this case, "sujeitas" [submitted]. From this concordance, we got information regarding:

(1) as rolhas são sujeitas a processos ou operações [stoppers are submitted to processes and operations]

Further analysing the Word sketch, some prepositional structures deserved our attention, such as the KWIC along with the prepositions “para” [for]; “de” [of]; “em” [in]; and “com” [with]. Some examples of concordances are presented below:

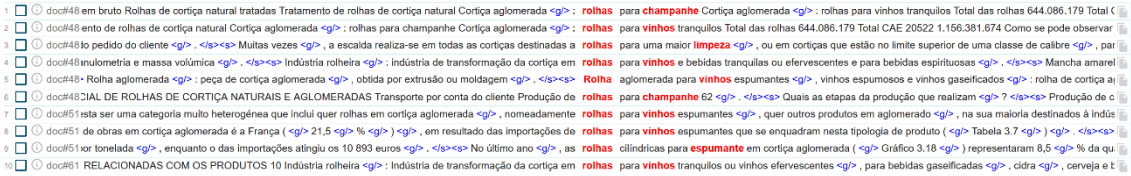


Figure 13: Concordance of rolha para + N (69 occurrences).

From the concordance shown above in Figure 13, we have identified 69 occurrences of the prepositional structure “rolha para + Noun” [stopper for + Noun], which has provided us with the information that stoppers have different functions. We enumerated this observation as a second piece of information:

(2) rolhas para champanhe / vinhos tranquilos/ vinhos efervescentes [stoppers for champagne / still wines / sparkling wines]

Looking at prepositional structures, we could observe that some polylexical terms either occur with the proposition “de” [of] or “em” [in]:

(i) “rolha de cortiça” [cork stopper] (1,013 occurrences)

(ii) “rolha em cortiça” [cork stopper] (161 occurrences)

The relevance of this morphosyntactic aspect will be further addressed in Section 3.4.1. And finally:

(iii) “rolha com + Noun” [stopper with + Noun] (158 occurrences), from which we obtained the piece of information that stoppers may have parts, namely “corpo” [body] or “cabeça” [head], or even defects, like “caleira”, as shown below, in Figure 14:

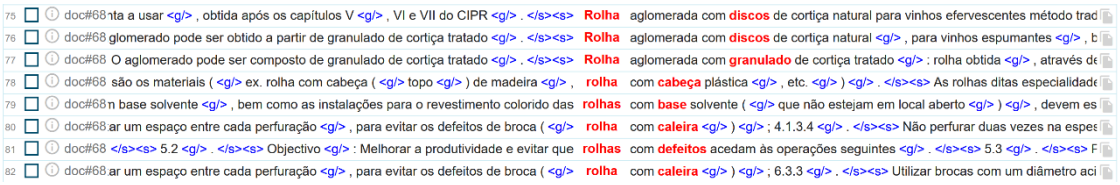


Figure 14: Concordance of “rolha com + N” [stopper with + N].

Concerning verbs co-occurring with the keyword, the verb “ser / estar” [to be] is the most quantitatively relevant verb in context and has called our attention given our awareness of its common predicative feature of introducing definitional contexts. However, the large scope of the Word sketch introduces some noise in the results, for most of the inflexions of the verb “ser/estar” [to be] occur with different purposes, and not only to define a given term, as we can see below in Figure 15.

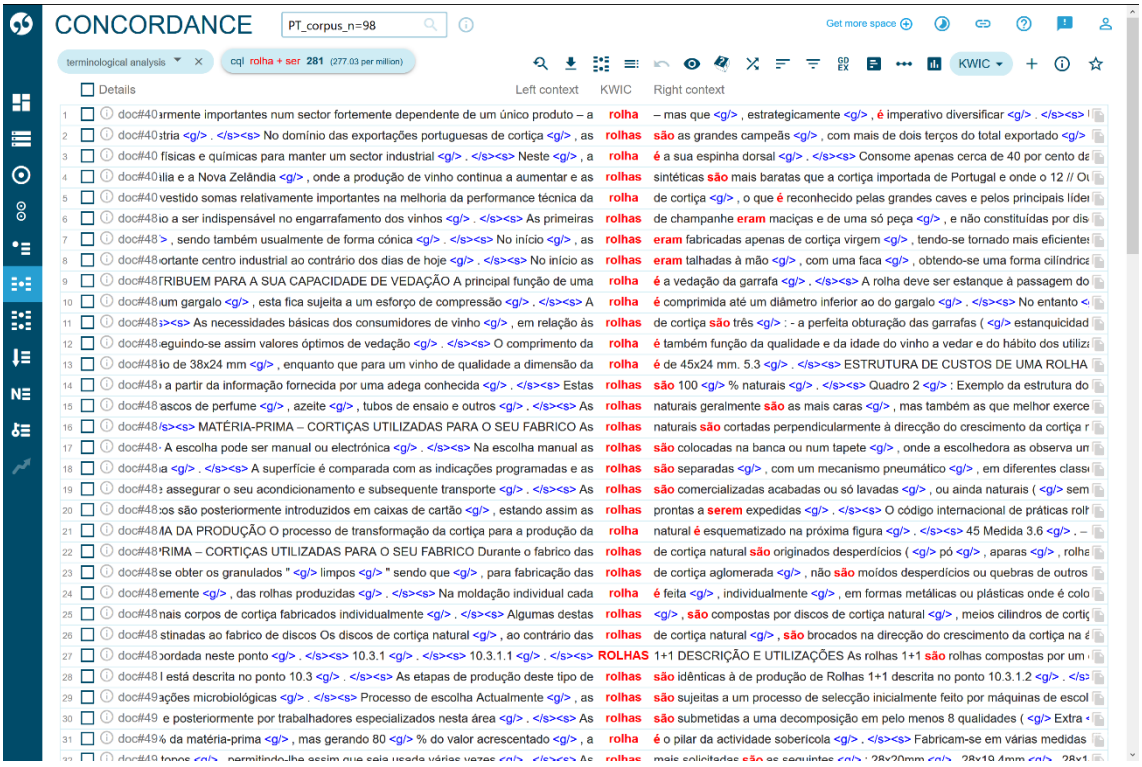


Figure 15: Concordance of “rolha + ser” [stopper + to be].

Figure 15 corresponds to the first page of the concordance of “rolha” + verb = “ser” [stopper + verb = to be], with 281 occurrences. From this concordance, we were able to extract a few contextual definitions and/or descriptions of the concepts within the scope of activities in the process of manufacturing cork stoppers; as well as a few contextual definitions and/or definitions for the keyword “rolha” [stopper], but mostly terms designating stoppers that were submitted to a given operation.

The contextual definitions we obtained from this concordance are systematised below in Table 12. It must be noted that most of these texts were retrieved from the examination of the distant context of the keyword, instead of the immediate left- or right-hand side of the KWIC. This means that when we identified a term in the concordance, we opened the text proper – by means of the tool – and retrieved additional terms from the surrounding context, as demonstrated below in Figure 16:

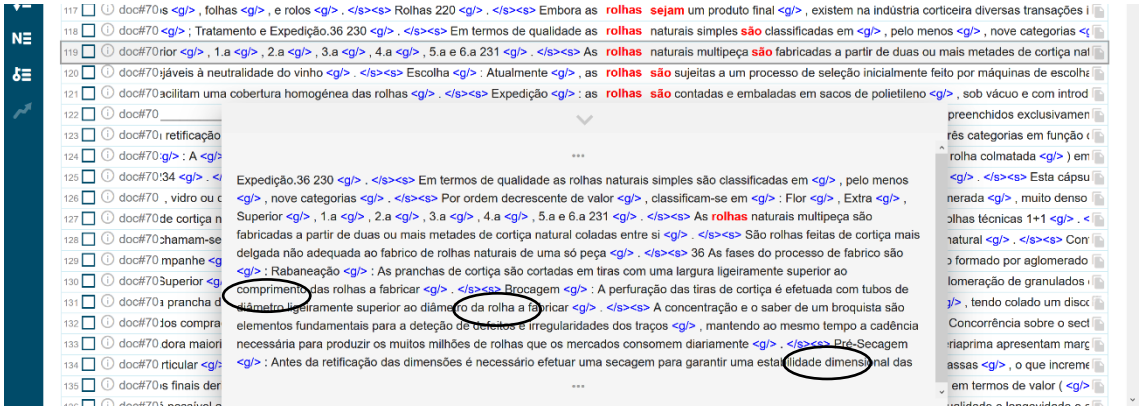


Figure 16: Several terms captured in the surrounding context of the keyword "rolhas".

As we can see in Figure 16, we managed to identify on line number 119 the term "rolhas naturais multipeça" [multi-piece natural stoppers] after the keyword search, namely "rolhas" [stoppers] highlighted in red, as well as its contextual definition – which can be observed immediately after the keyword on the flow of the sentence. Additionally, we were able to identify other candidate terms given the presence of descriptions within the scope of the manufacturing process as highlighted with a circle. One relevant aspect is to observe how the expert usually writes descriptions: in this case, before the description, the term is written with a capital letter and followed by a colon ":". These orthographic details are important for corpus advanced search, as we will further address.

Table 12: Contextual definitions captured via sketch word with "rolha" [stopper] as keyword.

No.	Term	definition /definitional context (pt)	literal translation (en)	File#
1		As **rolhas 1+1** são rolhas compostas por um corpo de cortiça aglomerada com dois discos de cortiça natural colados um em cada topo.	**1 + 1 stoppers** are stoppers composed of an agglomerated cork body with two natural cork disks glued together at each top.	27

2	**Rolhas de cortiça aglomerada:** **[Agglomerated cork stopper]**	rolhas feitas de cortiça granulada, constituídas de derivados da manufatura de rolhas naturais.	stoppers made of granulated cork, derived from the manufacture of natural stoppers.	69
3	**Rolha de cortiça capsulada:** **[Capsulated cork stopper]**	é uma rolha formada por um corpo de cortiça e uma cápsula em outro material.	it is a stopper formed by a body of cork and a capsule in another material.	69
4		As **rolhas naturais multipeça** são fabricadas a partir de duas ou mais metades de cortiça natural coladas entre si. São rolhas feitas de cortiça mais delgada não adequada ao fabrico de rolhas naturais de uma só peça.	**Multi-piece natural stoppers** are manufactured from two or more natural cork halves glued together. They are stoppers made from thinner cork not suitable for the manufacture of natural one-piece corks.	70
5		As **rolhas colmatadas** são rolhas de cortiça natural com os poros preenchidos exclusivamente com pó de cortiça resultante da retificação das rolhas naturais.	**Colmated stoppers** are natural cork stoppers with pores filled exclusively with cork powder resulting from the rectification of natural stoppers.	70
6		A **rolha capsulada** é uma rolha de cortiça natural (ou uma rolha colmatada) em cujo topo é colocada uma cápsula. Esta cápsula pode ser de madeira, PVC, porcelana, metal, vidro ou outros materiais.	The **capped stopper** is a natural cork stopper (or a colmated stopper) on the top of which a capsule is placed. This capsule can be made of wood, PVC, porcelain, metal, glass or other materials.	70
7		As **rolhas técnicas** são constituídas por um corpo de cortiça aglomerada, muito denso, com discos de cortiça natural colados no seu topo – ou em ambos os topos. As rolhas técnicas com um disco em cada topo são designadas rolhas técnicas 1+1. Com dois discos de cortiça natural em cada topo chamam-se rolhas técnicas 2+2, e com dois discos em apenas um dos topos chamam-se rolhas técnicas 2+0.	**Technical stoppers** are composed of a very dense agglomerated cork body, with natural cork discs glued to the top - or both tops. Technical stoppers with a disc on each top are called technical stoppers 1 + 1. With two discs of natural cork on each top they are called 2 + 2 technical stoppers, and with two discs on only one of the tops they are called 2 + 0 technical stoppers.	70
8		As **rolhas aglomeradas** são inteiramente fabricadas a partir da aglomeração de granulados da cortiça proveniente de subprodutos resultantes da produção de rolhas naturais.	**Agglomerated stoppers** are manufactured entirely from the agglomeration of cork granules from sub-products resulting from the production of natural corks.	70
9		As **rolhas de cortiça natural** são fabricadas por brocagem a partir de uma peça única de cortiça. Existem em forma cilíndrica ou cónica e em várias dimensões.	**Natural cork stoppers** are manufactured by drilling from a single piece of cork. They exist in cylindrical or conical form and in various dimensions.	93
10	**Rolhas naturais multipeça** **[Multi-piece natural stopper]**	As rolhas naturais multipeça são fabricadas a partir de duas ou mais peças de cortiça natural coladas entre si através de uma cola aprovada para estar em contacto com alimentos.	Multi-piece natural stoppers are manufactured from two or more pieces of natural cork glued together using an approved glue to be in contact with food.	93
11	**Rolhas naturais colmatadas**	As rolhas colmatadas são rolhas de cortiça natural com os poros (lenticelas) preenchidos	Colmated stoppers are natural cork stoppers with pores (lenticels) filled exclusively with cork powder	93

	[Colmated natural stopper]	exclusivamente com pó de cortiça resultante da rectificação das rolhas naturais.	resulting from the rectification of natural corks.	
12		As **rolhas microgranuladas** são rolhas com um corpo de cortiça aglomerada de grânulos finos, com dimensão média aproximada de 1 mm. Estes grânulos são colados entre si através de um adesivo aprovado para contacto alimentar.	The **micro-granulated stoppers** are stoppers with a fine granulated agglomerated cork body, with an average size of approximately 1 mm. These granules are glued together using an approved food contact adhesive.	93
13	**ROLHA:** **[STOPPER]**	Peça de cortiça, em geral cilíndrica, tronco-cónica ou prismática quadrangular, por vezes de arestas laterais boleadas ou chanfradas, constituída por um ou vários elementos colados e destinada a vedar os recipientes ou a contribuir para a sua estanquicidade.	Piece of cork, generally cylindrical, conical-trunk or square prismatic, sometimes with rounded or bevelled side edges, consisting of one or more glued elements and intended to seal containers or contribute to their tightness.	96
14	**ROLHA BOLEADA:** **[Rounded stopper]**	Rolha cujas arestas de um dos topos foram arredondadas por abrasão.	Stopper whose edges on one end were rounded by abrasion.	96
15	**ROLHA CHANFRADA:** **[Chamfered stopper]**	Rolha cujas arestas de um ou dois topos foram biseladas.	Stopper whose edges of one or two tops were bevelled.	96
16	**ROLHA COLADA (OU GEMINADA):** **[Glued stopper (or twined)]**	Rolha constituída por duas ou mais peças de cortiça coladas.	Stopper consisting of two or more pieces of cork glued.	96
17	**ROLHA COLMATADA:** **[Colmated stopper]**	Rolha submetida a um processo de colmatagem com pó de cortiça e colas, visando melhorar o seu aspecto visual.	Cork stopper submitted to a sealing process with cork powder and glues, aiming to improve its visual aspect.	96
18	**ROLHA DE CÁPSULA (OU DE CABEÇA):** **[Stopper with capsule (or with head)]**	rolha de flange em que esta é constituída por material diferente da cortiça.	stopper with a head consisting of a material other than cork.	96
19	**ROLHA DE FLANGE (OU CHAPÉU):** **[Stopper with hat]**	Rolha com corpo em forma cilíndrica ou cónica solidariamente encimado por um cilindro de maior diâmetro (chapéu).	Stopper with a cylindrical or conical body solidly topped by a larger diameter cylinder (hat).	96
20	**ROLHA DE IMITAÇÃO:** **[Simulated stopper]**	Rolha com forma de prisma quadrangular recto, de arestas laterais boleadas.	Stopper with a square prism shape, with rounded side edges.	96
21	**ROLHA LAVADA:** **[Washed stopper]**	Rolha que foi submetida a um tratamento químico com o objectivo de desinfectar e/ou homogeneizar a cor e/ou branquear.	Stopper that was submitted to chemical treatment with the aim of disinfecting and / or homogenizing the colour and / or bleaching.	96
22	**ROLHA MARCADA:** **[Marked stopper]**	Rolha cuja superfície lateral ou topos foram marcados a tinta ou a fogo.	Stopper whose side surface or tops have been ink or fire marked.	96
23	**ROLHA MISTA:** **[Mixt stopper]**	Rolha obtida por associação de peças em cortiça natural com peças em cortiça aglomerada.	Stopper obtained by associating pieces of natural cork with pieces of agglomerated cork.	96
24	**ROLHA PONÇADA:** **[Side surface sanded stopper]**	Rolha cuja superfície lateral foi submetida a uma operação de abrasão para a tornar cilíndrica ou diminuir o seu diâmetro.	Stopper whose side surface was submitted to an abrasion operation to make it cylindrical or to reduce its diameter.	96

Table 12, above, is our first systematisation of contextual definitions – i.e., definitions found in context – captured via an automatic co-occurrence search, namely

by means of the Word sketch feature, in which we parametrised “rolha” [stopper] as the keyword. As we can see, out of 24 contexts, eight of them do not have a term introducing the text definition, like for instance, context number 1. The examination of these texts led us to conclude that most of them are descriptions of the manufacturing process from which the stopper is obtained and/or point at the compositionality of the stopper – a description of the parts that compose the stopper. Hence, our designation of definitional contexts, above in Table 12. As for the textual definitions with an introductory term, most of these did not totally fulfil our expectations regarding the capture of polylexical terms designating different substance-type stoppers, unless for terms designating stoppers submitted to operations and/or with exotic shapes, as the ones shown in Table 12, fully capitalised (from 14- 24).

We could also observe in the Word sketch of “rolha” [stopper] that this term recursively occurs with Adjectives. The co-occurrences with the highest scores are listed below:

“rolha” + ADJ (pt)	ADJ + “stopper” (en)	Occurrences
rolha natural	natural stopper	114
rolha técnica	technical stopper	63
rolha capsulada	capsulated stopper	54
rolha cilíndrica	cylindric stopper	54
rolha aglomerada	agglomerated stopper	49

As mentioned before, this morphosyntactic structure, namely Noun + ADJ is expectable considering that the domain of cork, and particularly the manufacturing process of cork stoppers, involves activities of transformation. These activities are verbalised in discourse through verbs and the results of these activities are commonly qualities attributed to the recipient of the action, which in this case is the object designated by “rolha” [stopper].

We have subsequently searched for those 5 polylexical terms above, separately. However, when doing so, we realised that the results were not tuned with the ones provided by the Word sketch. For instance, the co-occurrence of “rolha” [stopper] and “técnica” [technical] has 63 occurrences within the results of Word sketch, while in a

separate simple concordance search, i.e., rolha* técnica* [96], this term has 119 occurrences.

Consequently, we decided to improve our search for terms and definitions by means of advanced searches, namely through regular expressions (REGEX), so we could capture in a more effective and summarised way knowledge rich contexts (KRC) – in the sense of Meyer (2001). The underlying goal of these KRC is capturing linguistic expressions pointing at and/or relating terms, for such contexts are commonly clues to infer the concepts designated by those terms to the extent that terminologists are able to grasp experts' knowledge through the observation of recurrent linguistic patterns used in discourse. By recurrent patterns, we mean that experts tend to recursively utter linguistic expressions that commonly relate terms, which in turn point at concepts, such as for instance, in the contextual definition number 21 and 24, as shown below (extracted from Table 12, p.100):

> 21. **Rolha** que foi submetida a um **tratamento químico** com o objectivo de desinfectar e/ou homogeneizar a cor e/ou branquear**.**
>
> **Stopper** that was submitted to **chemical treatment** with the aim of disinfecting and/or homogenising the colour and/or bleaching.

> 24. **Rolha** cuja superfície lateral foi submetida a uma **operação de abrasão** para a tornar cilíndrica ou diminuir o seu diâmetro.
>
> **Stopper** whose side surface was submitted to an **abrasion operation** to make it cylindrical or to reduce its diameter.

The underlined text is what we mean by recurrent patterns: "foi submetida a" [was submitted to]" is a linguistic expression that relates the term "rolha" [stopper] with a candidate term "operação de abrasão" [abrasion operation], in such a way that one can infer domain-specific information; in this case, we obtained information regarding treatments/operations.

[96] An asterisk is an operator that works as a wild card and "stands for zero or more occurrences of the preceding character." See: https://www.sketchengine.eu/guide/regular-expressions/

To those linguistic expressions pointing at knowledge rich contexts we would rather call linguistic markers – a topic that we will address in more detail in Section 4 – Linguistic analysis.

Based on the patterns we have captured within the definitional contexts, systematised above in Table 12 (p. 100), in addition to some terms referred to – such as "rolha de cortiça natural" [natural cork stopper] in both contexts number 5 and 6 – we have elaborated queries resorting to REGEX[97] – where a collection of symbols (also known as operators) can be used for pattern search – a syntax used in Corpus Query Language (CQL)[98] in Sketch Engine. This topic is further addressed in the next section.

3.4.1. Querying the corpus with CQL

As mentioned above, the FreeLing tagger has certain limitations, as does Sketch Engine itself. FreeLing cannot distinguish between adjectives and past participles, which is, as regards terminological work, highly limiting as observed in Costa (2001): in terms of probability, in Portuguese, the past participle is not usually part of the morphosyntactic structure of a term, while an adjective can be. This study was performed in the domain of Remote Sensing (Costa, 2001), where that characteristic was observed in the usage of the adjective "colorido" [colourful] as opposed to the usage of the past participle "colorido" [coloured]. This fact causes some noise in the results obtained from CQL queries. On the other hand, Sketch Engine does not allow semantic tagging, which would be a definite plus to retain certain types of forms while rejecting others thus contributing to the reduction of noise in the results obtained.

[97] REGEX "are used in CQL to specify patterns for values, e.g., [word = "dis.*"] [tag = "V.*"] finds words beginning dis- followed by a verb; [tag="J.*"] [word="[[:upper:]]*"] finds adjectives followed by an acronym (=word in capitals)." See https://www.sketchengine.eu/guide/regular-expressions/#toggle-id-2

[98] "The Corpus Query Language is a special code or query language used in Sketch Engine to search for complex grammatical or lexical patterns or to use search criteria which cannot be set using the standard user interface." See https://www.sketchengine.eu/documentation/corpus-querying/

Considering the tagger's limitations, we chose to elaborate simple REGEX as a start. The syntax of the simple queries was initially built with generic part-of-speech (POS) tags, as the ones shown below in Table 13:

Table 13: Most common POS Freeling tags used in our study

Grammatical category	**orthographic options**	**Generic POS tags (Freeling)**	**Specific POS tags**		**Character class**
noun		N	N.FS	:noun + feminine + singular	
verb		V	VM V.P	:verb + main; :verb + Past Participle	
adjective		A			
preposition		S			
determiner		D			
	punctuation	F	Fd	:colon	[[:punct:]]
	upper case				[[:upper:]]

There are also a few specific POS tags, in Table 13. These tags will be used at a second stage, after the following first simple query:

(1) [word="[[:upper:]]*"][tag="Fd.*"] 614 hits

CQL 1 aims at capturing textual data that match the structure:

ANY word written in upper case, followed by a form whose POS tag corresponds to a colon

The rationale behind our decision to search for such structure, as a start, is based on one of our observations after the texts systematised in Table 12 (p. 100), namely some textual definitions have terms written in upper case, followed by a colon, and then by the text that constitutes the definition proper. The results of CQL 1 were interesting, since we got 614 positive matches – to which we call "hits" – but obviously out of the range of "rolha" [stopper] given the large scope of the search, i.e., the character class [[:upper:]] is a regex that captures every form in upper case.

We decided thus to restrain CQL 1. This means that a filter was added to what was initially stated in CQL 1: instead of searching any words in upper case followed by a

colon, this time we have restrained the word to the lemma[99] “rolha” [stopper], which means that we are searching for any form of “rolha”[stopper].

(2) [word="[[:upper:]]*"&(lemma="rolha.*")][tag="Fd.*"] 21 hits

The 21 hits of CQL 2 clearly demonstrates that this query is too narrow, considering that the only word it aims at, given the structure of the regex, is:

the lemma of the word “rolha” [stopper] in upper case, followed by a colon

and nothing else. Among the results, we were able to capture again the definition for “ROLHA” [STOPPER], as already pointed out under number 13, above in Table 12 – our first systematisation of contextual definitions extracted by means of the Sketch word. This definition is, in our view, a good definition for the generic concept designated by “rolha” [stopper].

We decided to evolve CQL 2 inasmuch as we could capture the word “rolha” [stopper] but in a polylexical structure.

Recalling the terms systematised in Table 12 (p. 100), as well as the terms in the definitional contexts, some of those have a morphologic structure composed of four linguistic forms, e.g., “rolha de cortiça aglomerada” [agglomerated cork stopper]. Moreover, some of the terms have punctuation – i.e., curved brackets – which is also a form[100] for the tool. Therefore, the evolution of CLQ 2 has to do with the number of forms after the lemma of the word “rolha” [stopper]:

(3) [word="[[:upper:]]*"&(lemma="rolha.*")][]{0,6}[tag="Fd.*"] hits 208

CQL 3 aims at finding linguistic structures that match:

[99] According to Baker, et al.: “The canonical form of a word (the correct Greek plural is lemmata, although some people write the plural as lemmas and may consider lemmata to be somewhat pedantic). […] Lemmatised forms are sometimes written as small capitals, for example the verb lemma walk consists of the words walk, walked, walking and walks. In corpus studies, word frequencies are sometimes calculated on lemmata rather than types; words can also be given a form of annotation known as lemmatisation.” (2006, pp. 103-104).

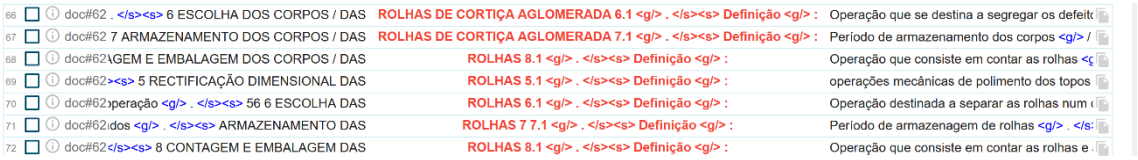
[100] A form is any character or string of characters between two white spaces in the text.

the lemma “rolha” in upper case, followed by ANY form – up to 6 forms – and finally followed by a colon

The results were quantitatively satisfactory, considering the 208 hits. From the concordance obtained from CQL 3, we were able to capture the terms listed below:

ROLHA DE CORTIÇA NATURAIS E AGLOMERADAS	[agglomerated and natural cork stopper]
ROLHAS	[stoppers]
ROLHAS DE CORTIÇA	[cork stoppers]
ROLHAS DE CORTIÇA NATURAL	[natural cork stoppers]
ROLHAS DE CORTIÇA AGLOMERADA	[agglomerated cork stoppers]
ROLHAS LAVADAS	[washed stoppers]
ROLHA	**[stopper]**
ROLHA BOLEADA	**[rounded stopper]**
ROLHA CHANFRADA	**[chamfered stopper]**
ROLHA COLADA	**[glued stopper]**
ROLHA COLMATADA	**[colmated stopper]**
ROLHA DE FLANGE	**[stopper with hat]**
ROLHA DE IMITAÇÃO	**[simulated stopper]**
ROLHA LAVADA	**[washed stopper]**
ROLHA MARCADA	**[marked stopper]**
ROLHA MISTA	**[mixt stopper]**
ROLHA PONÇADA	**[side surface sanded stopper]**
ROLHAS REJEITADAS	**[rejected stopper]**
ROLHA TOPEJADA	**[top polished stopper]**

Concerning the capture of definitions of the types of cork stoppers listed above, we obtained the same or the remaining definitions regarding cork stoppers that were submitted to a treatment/operation, identical to the ones we had initially systematised in Table 12. These terms are highlighted in bold in the list above. The absence of novel captures is a consequence of the filter used to restrain the matches for forms with upper cases: the search got circumscribed to two documents, given (1) the idiosyncrasies of a particular author and (2) the titles of paragraphs in a standard, regarding the activities and operations in the process of manufacturing stoppers. These activities are concepts partially designated by the terms we did not highlighted in bold in the list above. This means that, in addition to the KWIC matched by CQL 3, the concept defined is designated by a longer form, which we hypothesise as a candidate term whose polylexical form starts further away on the left-hand side of the concordance, as we can see below:



66	doc#62 . </s><s> 6 ESCOLHA DOS CORPOS / DAS	ROLHAS DE CORTIÇA AGLOMERADA 6.1 <g/> . </s><s> Definição <g/> :	Operação que se destina a segregar os defeit
67	doc#62 7 ARMAZENAMENTO DOS CORPOS / DAS	ROLHAS DE CORTIÇA AGLOMERADA 7.1 <g/> . </s><s> Definição <g/> :	Período de armazenamento dos corpos <g/> /
68	doc#62\GEM E EMBALAGEM DOS CORPOS / DAS	ROLHAS 8.1 <g/> . </s><s> Definição <g/> :	Operação que consiste em contar as rolhas <g
69	doc#62><s> 5 RECTIFICAÇÃO DIMENSIONAL DAS	ROLHAS 5.1 <g/> . </s><s> Definição <g/> :	operações mecânicas de polimento dos topos
70	doc#62\peração <g/> . </s><s> 56 6 ESCOLHA DAS	ROLHAS 6.1 <g/> . </s><s> Definição <g/> :	Operação destinada a separar as rolhas num (
71	doc#62\dos <g/> . </s><s> ARMAZENAMENTO DAS	ROLHAS 7 7.1 <g/> . </s><s> Definição <g/> :	Período de armazenagem de rolhas <g/> . </s:
72	doc#62</s><s> 8 CONTAGEM E EMBALAGEM DAS	ROLHAS 8.1 <g/> . </s><s> Definição <g/> :	Operação que consiste em contar as rolhas e

Figure 17: part of the concordance obtained after the CQL3.

Figure 17 depicts a part of the first page (out of three) of the results that match the regex of CQL 3. The KWIC – highlighted in red – is the last part of the terms that designate concepts of activities/operations, and the definition of the latter follows the colon. Although we are dealing with definitions of activities, and not the definitions of cork stoppers themselves, these definitions were retained in a data base, along with any other definition we could capture regarding cork stoppers, for they help us to understand and/or organise the domain under study. Moreover, within the structure of the terms designating activities, we were able to capture different terms, e.g., "ROLHAS DE CORTIÇA NATURAL" [natural cork stoppers].

So far, we have observed that terms designating cork stoppers that were submitted to a treatment/operation commonly have a polylexical structure within which the morphologic structure is Noun + Adjective, e.g.:

rolha NOUN colmatada ADJECTIVE [colmated stopper]

As far as we could observe, stoppers submitted to an operation are designated according to the involved operation. In the example above, the "rolha" [stopper] was submitted to the operation of sealing, which in Portuguese is designated as "colmatagem" or "colmatação" [sealing]. Thus, "rolha" [stopper] + "colmatagem" [sealing] = "rolha colmatada" [colmated stopper] [101]. Which leads us to assume that the adjective "colmatada" [sealed], within the morphologic structure of N + ADJ, derives from the past participle of the verb "colmatar" [to seal].

[101] Despite the inexistence of the adjective "colmated" in English, we have found the term "colmated corks" used as an equivalent for "rolha colmatada" in texts produced by native English speakers (see Taber, 2009).

Furthermore, within the text of contextual definitions of cork stoppers – independently of the focus of the definition, namely the substance of which it is made of or the treatment that intervened during its manufacture – the expert recursively uses the past participle (VPP) to describe, for instance, how the cork stopper was manufactured or obtained from, or its composition, as mentioned before. Some examples are listed below:

rolha **[feita]**$_{VPP}$ de	[stopper made of]
rolhas **[composta]** $_{VPP}$ por	[stopper composed of]
rolha **[obtida]** $_{VPP}$ por	[stopper obtained by]
rolha **[submetida]** $_{VPP}$ a	[stopper submitted to]

In light of these observations, we decided to create a new CQL, in order to capture linguistic expressions whose morphosyntactic structures match the pattern:

> ONLY the form “rolha” [stopper], followed by another form BUT whose grammatical category is either a Past Participle OR an Adjective

Besides the regex, we had to parametrise the advanced search to the default attribute “word”, so no other form but “rolha” – in lower case – would be matched:

(4) "rolha"[(tag="V.P.*")|(tag="A.*")] 148 hits

CQL 4 was quantitatively productive as regards the capture of terms morphologically composed of two forms, in which terms like “rolha técnica” [technical stopper], “rolha natural” [natural stopper], among others, could be retrieved given the (tag="A.* ") included in the main regex. We can see these terms below, in Figure 18 – a partial view of the concordance obtained with CQL 4.

109	doc#68 ıa broca <g/> . </s><s> 4.1.2 <g/> . </s><s> Objectivo <g/> : Obter uma	rolha cilíndrica	sem deformação nos limites dimensionais prescritos <g/> . </s><s> 4.1.
110	doc#68 /s><s> Definição <g/> : Operação de corte dos quadros para obter uma	rolha cilíndrica	<g/> . </s><s> 4.2.2.2 <g/> . </s><s> Objectivo <g/> : Obter uma rolha c
111	doc#68 indrica <g/> . </s><s> 4.2.2.2 <g/> . </s><s> Objectivo <g/> : Obter uma	rolha cilíndrica	sem deformações e em conformidade com os limites dimensionais pres
112	doc#68 ıma broca <g/> . </s><s> 6.2 <g/> . </s><s> Objectivo <g/> : Obter uma	rolha cilíndrica	sem deformação nos limites dimensionais prescritos <g/> . </s><s> 6.3
113	doc#68 hanframento <g/> , boleamento e <g/> / <g/> ou ponçagem do corpo da	rolha capsulada	<g/> . </s><s> 6.1.2 Evitar as superfícies facetadas <g/> , assegurando
114	doc#69 rolhas naturais <g/> . </s><s> Rolha de cortiça capsulada <g/> : é uma	rolha formada	por um corpo de cortiça e uma cápsula em outro material <g/> . </s><s>
115	doc#70 ıge <g/>) <g/> . </s><s> 32 Refira-se que a rolha de champanhe é uma	rolha técnica	<g/> , contudo será feita referência específica à rolha de champanhe se
116	doc#70 /> , B <g/> , C ou I <g/> , II <g/> , III <g/> . </s><s> 234 <g/> . </s><s> A	rolha capsulada	é uma rolha de cortiça natural (<g/> ou uma rolha colmatada <g/>) em
117	doc#70 </s><s> A rolha capsulada é uma rolha de cortiça natural (<g/> ou uma	rolha colmatada	<g/>) em cujo topo é colocada uma cápsula <g/> . </s><s> Esta cápsul
118	doc#70 e rolhas técnicas 2+0 <g/> . </s><s> 236 <g/> . </s><s> A qualidade da	rolha técnica	é bastante mais homogénea do que a da rolha natural <g/> . </s><s> C
119	doc#70 ▸ A qualidade da rolha técnica é bastante mais homogénea do que a da	rolha natural	<g/> . </s><s> Contudo <g/> , o padrão visual dos discos de cortiça natu

Figure 18: Concordance of CQL 4

However, we decided to add a filter to CQL 4 given the presence of noise among the results – e.g., “rolha inadequado” [inadequate stopper], in which the grammatical category of the adjective is masculine given the subject of the sentence (further in the left context of the concordance), namely “diâmetro” [diameter] – and restrained it into:

(5) "rolha"[(tag="V.P.*SF")|(tag="A.*")] 165 hits

Here, the regex [(tag="V.P.*SF") is an evolution of the tag [(tag="V.P.*")] previously declared in CQL 4, and intends to capture:

the past participle of ANY verb, BUT restrained to the inflexion of the Feminine Singular (3rd) Person given the preceding form “rolha” [stopper], whose grammatical category, in Portuguese, is a Singular Feminine Noun

The 165 hits of CQL 5 were quantitatively and qualitatively productive concerning terms composed of two forms and also to capture a few definitions. But the concordance of the matched linguistic structures had still noisy results given the impossibility of restraining the adjective genre into feminine with a regex like [tag="A.F.*"] – in the same way it effectively worked for nouns, e.g., [tag="N.F.* "]. Consequently, we could not capture only feminine adjective forms and match the regency dictated by the first form of the pattern, namely “rolha”[stopper]. Apparently, despite its reference in the tagset[102] of Freeling, the tagger did not recognise such specific POS, for the answer was nil results.

Subsequently, we decided to narrow down the regex of CQL 5 into the following:

(6) "rolha"[tag="V.P.*SF"] 69 hits

CQL 6 intends to capture:

ONLY forms “rolha”[stopper] in lower case followed by ANY Past Participle ONLY in Singular and Feminine inflection

[102] The Portuguese FreeLing part-of-speech tagset is available online: https://www.sketchengine.eu/portuguese-freeling-part-of-speech-tagset/

The concordance obtained from this CQL was the least noisy regarding unwanted results, but some terms, namely those whose morphologic structure is N+ADJ went silent in this concordance. Nevertheless, we were able to clearly identify and extract textual definitions and/or definitional contexts with this short but precise regex.

We have systematised the results of CQL 6 below in Table 14. Some may seem identical, but the documents are all different. We believe that this is a consequence of having some standards composing the corpus of analysis simultaneously with their older versions, therefore some definitions are (almost) identical. Furthermore, we have observed that the same definition(s) is recursively used across different texts: we assume this is due to the authority of the definition's source, namely the author/institution, thus, an intertextuality is observed in this corpus.

Table 14: Concordance obtained with CQL 6

#	left-hand side context	KWIC = CQL 6	right-hand side context
1	Indústrias produtoras de granulado de cortiça e / ou de	rolha aglomerada	- 2 Indústrias produtoras de rolhas de champanhe -
2	DESCRIÇÃO E UTILIZAÇÕES A **rolha natural trata-se de uma**	rolha composta	unicamente por cortiça, resultante da brocagem
3	**rolha de cortiça aglomerada com discos de cortiça natural:**	rolha formada	por um corpo em cortiça aglomerada e um ou dois discos
4	DE CORTIÇA NATURAIS E AGLOMERADAS **Rolha multi-peças:**	rolha constituída	por peças em cortiça natural coladas
5	1. Actividade 2. Produção de	rolha aglomerada	Acabamento de rolhas Comércio de rolhas
6	vedação e uniformizar a sua apresentação. **Rolha acabada:**	rolha acabada	pronta a usar, obtida após os capítulos V e VI do CIPR
7	**Rolha aglomerada com discos de cortiça natural para vinhos efervescentes método tradicional:**	rolha formada	por um corpo de cortiça aglomerado, tendo um ou mais discos de cortiça colado num dos topos
8	**Rolha aglomerada com granulado de cortiça tratado:**	rolha obtida	através de um processo de moldagem
9	**Rolha de cortiça aglomerada inserida totalmente no gargalo com discos de cortiça natural para vinhos tranquilos e vinhos frisantes:**	rolha formada	por um corpo de cortiça aglomerada, tendo um ou mais discos
10	tratado. **Rolha de cortiça aglomerada por extrusão:**	rolha obtida	, através de um processo de extrusão
11	Versão 6.03 9 **Rolha de cortiça aglomerada por moldagem:**	rolha obtida	, através de um processo de moldagem
12	granulado compreendida entre 0,25 e 8 mm. **rolha multi-peças :**	rolha constituida	por várias peças em cortiça natural coladas
13	natural coladas entre si. **Rolha semi-acabada :**	rolha semi-manufacturada	transformada durante IV do CIPR
14	capítulo IV do CIPR. **Rolha semi-manufacturada:**	rolha obtida	após o capítulo III do CIPR. Rolha: produto obtido
15	. **Rolha de cortiça natural colmatadas** ISO 633 -	rolha feita	de cortiça natural. NOTA: As rolhas de cortiça

16	. **Rolha de cortiça aglomerada nova geração** ISO 633 –	rolha obtida	pela aglutinação de grânulos de cortiça com dimensão
17	vinho 7.1.1. Comprimento da rolha. O comprimento da	rolha seleccionada	deve estar de acordo com o nível de enchimento da garrafa
18	comprimento parcial Medida do corpo de cortiça de uma	rolha capsulada	6.2.1.3 diâmetro Maior distância entre dois pontos
18	das rolhas relacionadas com o acabamento 6.2.2.1	**rolha chanfrada**	Rolha cujas arestas de um ou dois topos foram biseladas
20	arestas de um ou dois topos foram biseladas 6.2.2.2	**rolha ponçada**	Rolha cuja superfície lateral foi submetido a uma rectificação dimensional
21	foi submetido a uma rectificação dimensional 6.2.2.3	**rolha boleada**	Rolha cujas arestas de um ou dois topos foram arredondadas
22	e cortiça natural colado num dos topos 6.3.6	**rolha capsulada**	Rolha em cortiça natural, natural colmatada
23	lenhificada, com uma sobre-espessura anormal 6.6.9	**rolha deformada**	Rolha que apresenta uma protuberância no corpo
24	Rolha que apresenta uma protuberância no corpo 6.6.10	**rolha biselada (assobio)**	Rolha que apresenta uma ou as duas extremidades enviesadas (oblíquas), devido a uma brocagem imperfeita
25	o corpo da rolha, provocado por brocagem de uma	rolha sobreposta	à anterior 6.6.12 rolha preguenta ou rolha lenhosa Rolha que
26	uniformizar a sua apresentação. **Rolha acabada:**	rolha acabada	pronta a usar, obtida após os capítulos V, VI e VII do CIPR
27	**Rolha aglomerada com discos de cortiça natural para vinhos efervescentes método tradicional:**	rolha formada	por um corpo de cortiça aglomerado, tendo um ou mais discos
28	**Rolha aglomerada com granulado de cortiça tratado:**	rolha obtida	, através de um processo de moldagem
29	**Rolha de cortiça aglomerada inserida totalmente no gargalo com discos de cortiça natural, para vinhos tranquilos e vinhos frisantes:**	rolha formada	por um corpo de cortiça aglomerada, tendo um ou mais discos de cortiça natural colado(s) num ou nos dois topos
30	tratado. **Rolha de cortiça aglomerada por extrusão:**	rolha obtida	, através de um processo de extrusão, por aglutinação
31	e 8 mm. **Rolha de cortiça aglomerada por moldagem:**	rolha obtida	, através de um processo de moldagem, por aglutinação
32	& Práticas Gerais Obrigatórias **Rolha multi-peças:**	rolha constituída	por várias peças em cortiça natural coladas entre si.
33	cortiça natural coladas entre si. **Rolha semi-acabada:**	rolha semi-manufacturada	transformada durante o capítulo IV do CIPR. **Rolha:** produto obtido de cortiça e/ou cortiça aglomerada constituído por uma ou mais peças, destinado a vedar garrafas ou outros recipientes e a preservar o seu conteúdo
34	capítulo IV do CIPR. **Rolha semi-manufacturada:**	rolha obtida	após o capítulo III do CIPR. Rolha: produto obtido
35	chanframento, boleamento e/ou ponçagem do corpo da	rolha capsulada	.6.1.2 Evitar as superfícies facetadas,
36	uniformizar a sua apresentação. **Rolha acabada:**	rolha acabada	pronta a usar, obtida após os capítulos V, VI e VII do CIPR
37	**Rolha aglomerada com discos de cortiça natural para vinhos efervescentes método tradicional:**	rolha formada	or um corpo de cortiça aglomerado, tendo um ou mais discos
38	**Rolha aglomerada com granulado de cortiça tratado:**	rolha obtida	, através de um processo de moldagem, por aglutinação
39	**Rolha de cortiça aglomerada inserida totalmente no gargalo com discos de**	rolha formada	por um corpo de cortiça aglomerada, tendo um ou mais discos

	cortiça natural, para vinhos tranquilos e vinhos frisantes:		
40	tratado. **Rolha de cortiça aglomerada por extrusão:**	rolha obtida	, através de um processo de extrusão, por aglutinação
41	obrigatórias **Rolha de cortiça aglomerada por moldagem:**	rolha obtida	, através de um processo de moldagem, por aglutinação
42	as naturais colmatadas". **Rolha multi-peças:**	rolha constituída	por várias peças em cortiça natural coladas entre si
43	natural coladas entre si. **Rolha semi-acabada:**	rolha semi-manufacturada	transformada durante o capítulo IV do CIPR.
44	apítulo IV do CIPR. **Rolha semi-manufacturada:**	rolha obtida	após o capítulo III do CIPR. Traço / rabanada
45	nframento, boleamento e/ou ponçagem do corpo da	rolha capsulada	.6.1.2 Evitar as superfícies facetadas, assegurando
46	lhas naturais. **Rolha de cortiça capsulada:** é uma	rolha formada	por um corpo de cortiça e uma cápsula em outro material
47	de colmatagem: A, B , C ou I , II , III . 234 . A	**rolha capsulada**	é uma rolha de cortiça natural (ou uma rolha colmatada)
48	capsulada é uma rolha de cortiça natural (ou uma	rolha colmatada	) em cujo topo é colocada uma cápsula.
49	. A **rolha de cortiça natural colmatada** é uma	rolha feita	de cortiça natural em que são obturadas as suas lenticelas
50	cortiça natural. **Rolha de cortiça natural colmatada –**	rolha feita	de cortiça natural em que são obturadas as lenticelas das rolhas e/ou dos discos da cortiça com uma mistura de colas e pó de cortiça proveniente dos acabamentos dimensionais das rolhas de cortiça natural
51	adesivo. **Rolha de cortiça aglomerada nova geração –**	rolha obtida	pela aglutinação de grânulos de cortiça com dimensão compreendida entre 0,25 mm e 8 mm
52	da peça é ligeiramente superior ao comprimento da	rolha pretendida	. E Formação das pilhas. Ainda na floresta
53	a largura é ligeiramente superior ao comprimento da	rolha pretendida	(NP 273). A obtenção das rolhas, opera
54	conjunto destas características. O preço de uma	rolha dita	de qualidade extra ou superior poderá ser dezenas de vezes mais e
55	Liège.34 05.05.8 - Rolhas Capsuladas A	**rolha capsulada**	é uma rolha de cortiça em cujo topo é colocada uma cápsula, de
56	idas (comprimento x diâmetro) mais comuns são: A	**rolha capsulada**	é geralmente utilizada em vinhos licorosos/ generosos ou em
57	e, como tal, extremamente raro de aparecer numa	rolha terminada	; • Defeitos de fabrico. São problemas que podem
58	para aumentar os benefícios de utilização de uma	rolha certificada	, poderão obter certificação de cadeia de custódia que lhes
59	mercado das **rolhas técnicas** (designadas como **"1+1"** –	rolha constituída	por um disco de cortiça natural em ambos os topos e um corpo de aglomerado de cortiça)
60	11:35 AM 05.05.8 - Rolhas Capsuladas A	**rolha capsulada**	é uma rolha de cortiça em cujo topo é colocada uma cápsula, de madeira, PVC , porcelana, metal, vidro ou outros materiais
61	didas (comprimento x diâmetro) mais comuns são: A	**rolha capsulada**	é geralmente utilizada em vinhos licorosos/ generosos ou em
62	extremamente raro de aparecer numa	rolha terminada	; Defeitos de fabrico. São problemas que
63	mercado das **rolhas técnicas** (designadas como **"1+1"** –	rolha constituída	por um disco de cortiça natural em ambos os topos e um corpo
64	as com um diâmetro maior que as rolhas normais.	**rolha aglomerada**	– rolhas com um corpo de cortiça aglomerada; rolha micro

65	grânulos finos, compreendido entre 0,25 mm e 8 mm;	**rolha capsulada**	– rolha de cortiça natural em cujo topo é colada uma cápsula de
66	aglomerado decorativo (alta frequência), blocos de	rolha aglomerada	(alta frequência), rolha aglomerada por extrusão e por
67	blocos de rolha aglomerada (alta frequência),	rolha aglomerada	por extrusão e por moldação. Resinas fenólicas -
68	ou caleira - sulco longitudinal na superfície da	rolha provocada	por se brocarem as rolhas muito juntas; Fenda ou racha - fissura
69	.Refira-se que muitos consumidores associam uma	rolha marcada	com um bonito desenho a um vinho de qualidade, e que esta pode

As systematised and highlighted in bold above in Table 14, we have retained some definitions among which some texts are what we consider as descriptions of the concept given the structure of the definitional text, such as the ones shown in lines 2, 47, 49, 55, 56, 60 and 61. In these texts, we can see that terms are not firstly enunciated, nor followed by the definition of the concept being designated. Instead, they start with an article – highlighted in red – before the term and continue a description either of the substance of which the stopper is made of or its constituent pieces, or even the function of the stopper, such as shown in line 61. Nevertheless, descriptions are still valid for our terminological work given the information they convey.

From the observations we have outlined so far, we believe to have demonstrated the reason for our decision of searching the corpus through advanced CQL, where the initial regex were elaborated with generic labels in the first place so that we could capture silences on the one hand, and then these regex progressively evolved into more restrained ones, on the other hand, so that we could avoid the noisy results caused by the regex with generic labels, i.e., labels without genre specification.

The iterative work to attain the above observations, namely the elaboration of regex and the back and forth of their evolution/involution, aims at facilitating clear results inasmuch as patterns are matched according to specific linguistic expressions that fulfil our expectations. In the case of the last CQL (6), the linguistic expressions identified in the first definitions we have initially systematised in Table 12 (p. 100) were the most productive patterns to take into account for regex elaboration, such as the Past Participle. The results of CQL 6, systematised above in Table 14, demonstrate how

particular linguistic patterns are knowledge rich contexts, for we managed to extract several definitions though the identification of such patterns.

Finally, and considering the satisfactory results of CQL 6, we decided to expand its regex and merge it with the results we have observed with the Word sketch of "rolha"[stopper], as follows:

(7) "rolha"[(tag="D.*")|(tag="S.*")]?[tag="A.*"]?"cortiça"?[]{0,4}"rolha"{0,4}[tag="V.P.*SF"]

CQL 7 intends to capture linguistic patterns with the exact sequence:

> the word "rolha"[stopper] followed by ANY Determiner OR ANY Preposition (or none of both); ANY Adjective (or not); the word "cortiça" [cork] (or not); ANY form (from zero to 4); the word "rolha" (from 0 to 4) ; and finally ANY Past Participle BUT Feminine Singular

We obtained a concordance with 167 hits with CQL 7. This CQL was very effective to capture both polylexical terms and definitions (see Annex 4). Among the 167 hits, we could identify 90 lines containing either a description or a definition, although some of these repeated – a consequence from the operator "?", which means zero or one occurrence of the previous form, therefore, duplicating some results.

Besides the definitions and/or terms already systematised above in Table 14 – the concordance of CQL 6 – we managed to extract with CQL 7 additional terms and definitions/descriptions, as systematised below in Table 15:

Table 15: Some terms and definitions/description captured with CQL 7

#	left-hand side context	KWIC = CQL 7	right-hand side context
1	**Rolha composta :**	rolha de cortiça aglomerada	, composta de, pelo menos, 51 % de granulado de cortiça (em peso), com uma granulometria de 0,5 mm (mínimo), peso específico máximo de 60 kg/m3 e um teor em água igual ou inferior a 8 % (Norma ISO 2190)
2	mercado das **rolhas técnicas (designadas como "1+1" –**	rolha constituída	por um disco de cortiça natural em ambos os topos e um corpo de aglomerado de cortiça
3	**N+N (Um mais Um ou Rolha** ISO 633 –	rolha com um corpo de cortiça aglomerada	e Técnica) n discos de cortiça natural colados num ou em ambos os topos. NOTA: Nesta designação n indica o número de discos usados.
4	e composta, pelo menos, por 51 % de granulado de cortiça, em peso. 6.3.2.1	**rolha de cortiça aglomerada tratada**	* Rolha obtida pela aglutinação de granulado de cortiça com dimensão compreendida entre 0,25mm e 8mm

5	cone 6.4.3 **rolha cilíndrico-cónica**	rolha com uma parte cilíndrica justaposta	a outra parte de forma cónica
6	cortiça natural coladas umas às outras.875 .	**rolha técnica (**	**também designada por ' Um mais Um')** é constituída por um corpo de cortiça aglomerada e 2 discos de cortiça natural colados num ou em ambos os topos
7	Rolha de micrograniulado. 919. A	**rolha técnica** para vinhos espumantes é produzida	a partir de um corpo formado por aglomerado de grânulos de cortiça
8	**Rolha N+N (Um mais Um ou Rolha Técnica)**	rolha com um corpo de cortiça aglomerada	e n discos de cortiça natural colados num ou em ambos os topos (Nota: Nesta designação n indica o número de discos usados
9	agrupar-se nas seguintes categorias	**rolha natural** – peça única, extraída	por brocagem de um traço de cortiça

Our main interest on these last definitions is tied with the term “rolha técnica” [technical stopper] given the several designations it may have, namely “N+N”, “rolha N+N” [N+N stopper], “1+1”, and “um mais um” [one plus one]. The definition of the object points at several discs glued on one or both tops of the stopper’s body. The letter “N” means the number of discs, thus, the possibility of 3 types of technical stoppers, namely 1+1, 2+2, and 0+2.

To conclude this section, we have elaborated one last CQL (8) to demonstrate how dynamic this corpus-search work is given the iterative tasks that are involved for the creation of regex for corpus exploration by means of text mining strategies. We must stress that regex evolve and/or involve depending on the observations of the results of each of those regexes.

(8) "rolha" [(tag="D.*")|(tag="S.*")]?"cortiça"[]{0,4}"rolha"[]{0,4}[tag="V.P.*SF"]

CQL 8 is parametrised to search lemmas by default, which means that forms written with “ ” are captured in all inflexions of the word. This CQL captures the following patterns, in the exact sequence:

the lemma of the word “rolha” [stopper], followed by ANY Determiner OR ANY Preposition (or none of both); the lemma of the word “cortiça” [cork]; ANY form (from zero to 4); the lemma of “rolha”; ANY form (from zero to 4); and finally, ANY Past Participle BUT Feminine Singular

As a result, we obtained a concordance with 26 hits, corresponding to 23 contextual definitions, as depicted below in Figure 19.

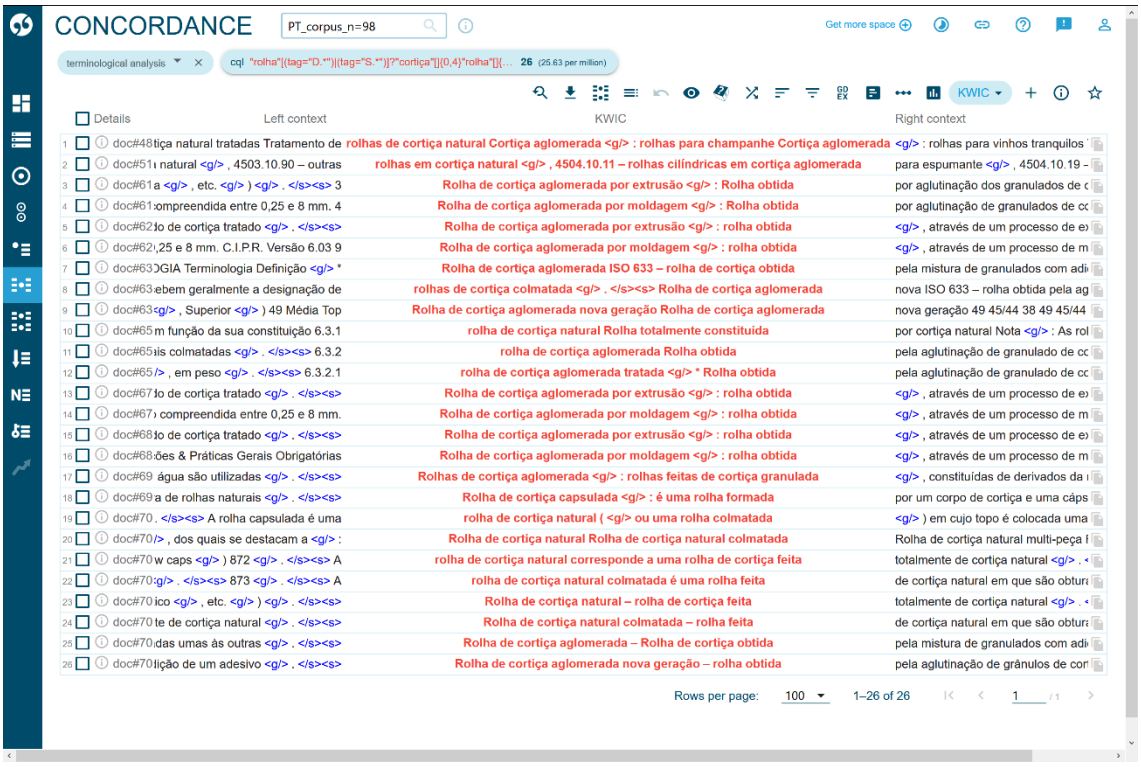


Figure 19: Concordance of CQL 8

As we can see in Figure 19, the tag we have used at the end of all regex is a Past Participle and effectively captured verbs pointing at either the underlying operation or the means to obtain the manufactured object. As mentioned before, we have based this search on the observations made on the first definitions extracted, in which the expert recursively uses that linguistic expression. This was one of the strategies we have used to capture contextual definitions, but many others are possible given the different linguistic structures used by the expert to describe the concept.

To conclude with, the ability to capture definitions straightforwardly is not attainable, nor is it possible to capture in one single regex all the possible linguistic expressions used in discourse to convey a piece of information or describe a given concept. Admitting the opposite would lead us to incur in the erroneous assumption that discourse is based on a rigid model, which contradicts the richness of language and the countless possible linguistic structures that speakers have to verbalise their conceptualisations.

3.5. Ten (10) definitions to organise a typology of cork stoppers

Among the whole set of descriptions or textual definitions we have semi-automatically extracted from the Cork Corpus pt, we decided to select ten (10) textual definitions to analyse linguistically and conceptually. The results of these analyses will be the basis for our terminological task of modelling the underlying knowledge.

To start with, for the knowledge organisation of the domain under focus, the selected ten definitions are quantitively suitable to build a micro domain-ontology, where a typology of cork stoppers will be defined through formal logic descriptions (see Section 6, p. 215). The ten definitions, which are originally written in Portuguese, are systematised below, in Table 16:

Table 16: Ten (10) definitions to organise a typology of cork stoppers

#	10 definitions (literal translations from pt)	10 definitions (pt) extracted from the Cork corpus
1	**stopper** Product obtained from natural cork and / or agglomerated cork, consisting of one or more pieces, intended to seal bottles or other containers and to preserve their contents. (5.1 - NORM)	**rolha** Produto obtido da cortiça natural e / ou de cortiça aglomerada, constituído por uma ou mais peças, destinado a vedar garrafas ou outros recipientes e a preservar o seu conteúdo. (5.1 - NORM)
2	**STOPPER** piece of cork, usually cylindrical, conical or prismatic quadrangular, sometimes with rounded or chamfered lateral edges, consisting of one or several glued elements and intended to seal the containers or contribute to their water tightness. (7.8 – TECH)	**ROLHA** peça de cortiça, em geral cilíndrica, troncocónica ou prismática quadrangular, por vezes de arestas laterais boleadas ou chanfradas, constituída por um ou vários elementos colados e destinada a vedar os recipientes ou a contribuir para a sua *estanquicidade[103] (7.8 – TECH)
3	**natural cork stopper** Stopper consisting entirely of natural cork Note: Natural cork stoppers that have been submitted to the sealing operation (see 6.5.5) are commonly referred to as colmated natural stoppers. (5.5 – NORM)	**rolha de cortiça natural** Rolha totalmente constituída por cortiça natural. Nota: As rolhas naturais que tenham sido submetidas à operação de colmatagem (ver 6.5.5) são comummente designadas por rolhas naturais colmatadas. (5.5 – NORM)
4	**colmated natural cork stopper** The colmated natural cork stopper is a stopper made of natural cork in which its lenticels are filled with a mixture of glues and	**rolha de cortiça natural colmatada** A rolha de cortiça natural colmatada é uma rolha feita de cortiça natural em que são obturadas as suas lenticelas com uma mistura de colas e pó de

[103] Spelling choice of the expert. This form was found in several occurrences (29) across different texts in the corpus of analysis although it does not exist in Portuguese dictionaries.

	cork powder from the dimensional finishing processes of natural cork stoppers. (6.1 – REP)	cortiça proveniente dos acabamentos dimensionais das rolhas de cortiça natural. (6.1 – REP)
5	**agglomerated cork stopper** Stopper obtained by the agglutination of cork granules with a size between 0,25 mm and 8 mm, with addition of binders, by means of extrusion or moulding and composed of at least 51% by weight of cork granules. (5.5 – NORM)	**rolha de cortiça aglomerada** Rolha obtida pela aglutinação de granulado de cortiça com dimensão compreendida entre 0,25mm e 8mm, com adição de ligantes, através de extrusão ou moldagem e composta, pelo menos, por 51 % de granulado de cortiça, em peso. (5.5 – NORM)
6	**agglomerated stopper:** piece of agglomerated cork, obtained by extrusion or moulding (3.1 – STUD)	**rolha aglomerada:** peça de cortiça aglomerada, obtida por extrusão ou moldagem (3.1 – STUD)
7	**n+n stopper** Stopper formed by a body of agglomerated cork and "n" disks of natural cork glued to one or both ends. N.B.: In this designation, "n" indicates the number of disks used. (5.5 – NORM)	**rolha n+n** Rolha formada por um corpo de cortiça aglomerada e "n" discos de cortiça natural colados num ou em ambos os topos. Nota: Nesta designação, "n" indica o número de discos utilizados. (5.5 – NORM)
8	**technical stopper** Technical stoppers are composed of a very dense body of agglomerated cork with disks of natural cork glued to one end - or to both ends. Technical stoppers with one disk on each end are called 1+1 technical stoppers; those with two disks of natural cork on each end are called 2+2 technical stopper; and those with two disks glued at only one of the ends are called 2+0 technical stoppers. (6.1 – REP)	**rolha técnica** As rolhas técnicas são constituídas por um corpo de cortiça aglomerada, muito denso, com discos de cortiça natural colados no seu topo – ou em ambos os topos. As rolhas técnicas com um disco em cada topo são designadas rolhas técnicas 1+1. Com dois discos de cortiça natural em cada topo chamam-se rolhas técnicas 2+2, e com dois discos em apenas um dos topos chamam-se rolhas técnicas 2+0. (6.1 – REP)
9	**rounded stopper** Stopper whose edges of one or two ends were rounded by abrasion. (5.5 – NORM)	**rolha boleada** Rolha cujas arestas de um ou dois topos foram arredondadas, por abrasão. (5.5 – NORM)
10	**marked stopper** Stopper whose lateral surface or ends were marked in ink or by fire (7.6 – TECH)	**ROLHA MARCADA** Rolha cuja superfície lateral ou topos foram marcados a tinta ou a fogo. (7.6 – TECH)

As we can see in Table 16 above, we have recorded the original ten textual definitions in Portuguese and their corresponding (literal) translations[104] in English. The linguistic expressions underlined in the textual definitions are the linguistic patterns we have observed that recursively occur in textual definitions. It is on these recursive linguistic expressions that our linguistic analysis will mainly focus, as further demonstrated during the analysis of Definitions 1, 2, 3 and 4, in the next section (4).

[104] The translation of all definitions is ours.

Furthermore, the linguistic expressions and underlying information are at the core of our linguistic analysis in order to infer a micro-concept system of the domain and finally feed a domain-ontology – the final task in this study (see Section 6).

The remaining textual definitions (from 5 to 10) were also analysed, both linguistically and conceptually to the extent that we could formally model a typology of cork stoppers; however, their analysis will not be demonstrated step-by-step, as we will address the first four definitions. We have chosen these particular 4 definitions given (1) the large scope of the generic term, (2) the information regarding the compositionality – i.e., the parts – and finally (3) the different types of substance, e.g., natural cork or agglomerated cork.

Linguistic analysis

4. Definition

Much has been said about *definition* – a matter that has been regarded as a classical subject after the ancient Greek philosophers. Socrates, Plato, and Aristotle are well known for their studies on this topic. Answering questions such as "What is this?" was seen as highly relevant, as pointed out by Smith:

> The definition (horos, horismos) was an important matter for Plato and for the Early Academy. Concern with answering the question "What is so-and-so?" are at the center of the majority of Plato's dialogues, some of which (most elaborately the Sophist) propound methods for finding definitions. External sources (sometimes the satirical remarks of comedians) also reflect this Academic concern with definitions. Aristotle himself traces the quest for definitions back to Socrates. (Smith, 2020)

Among the above-mentioned philosophers, Aristotle was the one that produced one of the most important works on this matter, particularly with regards to his logic premises to answer the question "what is so-and-so". For Aristotle, the notion of "what it is to be" is so pervasive that it becomes formulaic to such extent, that a definition is what expresses "what it is to be" or, in modern terminology, its essence (Smith, 2020).

According to Rey (1990), "*definition"* is a polysemic term. This author claims that different types of definitions are possible to envisage, such as those that have an ontological purpose, where the focus is to describe the essence of a given logical-linguistic operation needed to represent language signs – in the Saussurean sense – in a controlled manner. This is the Aristotelian type of definition, after his epistemological pursuit of "un discours des limites" (*ibid.*).

This need to bound knowledge is also conveyed by the sense of *limit* underlying the Latin word for definition when observing its morphosyntactic decomposition: *de-finitio,* the sense of finitude – delimitation – from the form *finitio* (*ibid.*). Meaning, that defining is an operation at the level of abstraction, in which the concept – a "unit of knowledge created by a unique combination of characteristics*"* (ISO/FDIS 1087, 2019 (E), p. 3) – is delimited by the conceptual relations established by differentiation. This

differentiation paves the way to knowledge organisation, where the term – a "verbal designation of a general concept in a specific subject field" (ISO 1087-1, 2000, p. 6) – occupying the position of *definiendum* in a given definitional text is the assigned element bridging the gap between what is from the level of abstraction and what is from the level of language.

In sum, there are as many types of definitions as there are purposes for the definitions, i.e., definitions may be philosophic in the metaphysical sense of Aristotle's discourse or philological, in the pragmatic sense of a social product within which the stipulative, constructive or descriptive procedures are observed. Our interest, however, falls under the type of definition that serves to differentiate a given concept from another one in a *concept system* (ISO/FDIS 1087, 2019 (E)) pertaining to a well-defined domain. Delimiting the domain is a task performed in the scope of terminological work, in which terminology – in the sense of term collection – mirrors a structured organisation. The structured organisation of the domain's terminology is the terminological work in itself, which explains the close link between term and definition. The interdependence among concept, term, domain, and definition is what constitutes the terminological triangle, as highlighted by De Bessé (1990, p. 251):

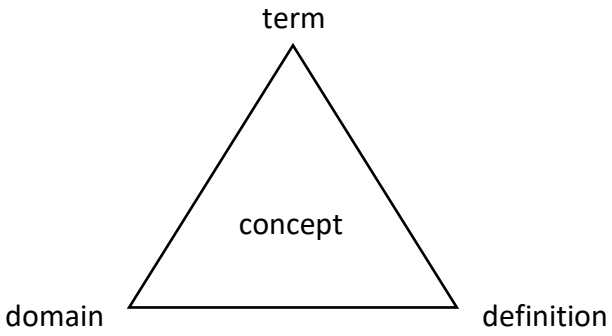


Figure 20: Terminological triangle

4.1. Intensional definition

The definition plays a major role in our study given its properties of objective information and/or inherent linguistic and conceptual representations. Actually, the latter is in accordance with the "operation" and "résultat" mentioned by Rey:

> Les mots définition et terme sont liés par un trait commun : ils désignent à l'origine l'assignation d'une limite, d'une fin (dé-finir) et son résultat (terme). Au plan notionnel, pour qu'un nom ait droit au titre de terme, il faut qu'il puisse, en tant qu'élément d'un ensemble (une terminologie), être distingué de tout autre. Le seul moyen pour exprimer ce système de distinctions réciproques est l'opération dite définition. (1979, p. 40)

Following the assertion of Rey, terms and definitions are at the core of the terminological work, where both the former and the latter ought to be unambiguously differentiated. For that purpose, we will address this differentiation of terms and definitions by describing the intension of concepts. According to ISO TC 37 1087, a "definition that conveys the intension of a concept by stating the immediate generic concept and the delimiting characteristic(s)" (ISO/FDIS 1087, 2019 (E), p. 7) is an ***intensional definition***.

According to ISO 704 (2009), the role of an intensional definition is to provide concise information without delivering too much information to the extent that one unambiguously recognises, at the level of abstraction, the place of a concept in the concept system by differentiating it from the other concepts. The structure of the information made explicit in the intensional definition leads to the recognition and differentiation of a given concept, i.e., when defining a given concept in natural language, the

> superordinate concept [is] immediately above, followed by the delimiting characteristic(s). The superordinate concept situates the concept in its proper context in the concept system (i.e. 'mice' among 'pointing devices', 'trees' among 'plants'). In practice, intensional definitions are preferable to other types of definitions and should

be used whenever possible as they most clearly reveal the characteristics of a concept within a concept system. (ISO 704, 2009, p. 22)

Given the brief information conveyed by the above-mentioned structure, intensional definitions are considered the most explicit and precise method of concept definition (*ibid.*). Hence, and based on this assumption, the structure of an intensional definition is the model on which we ground our linguistic analysis of the definitions found in the corpus of analysis, on the one hand, and propose new definitions, on the other.

4.1.1. Essential characteristics

Before we start describing our methodology for the systematisation of the concepts from the domain under study, there are some key terminology concepts we must address first and foremost, namely the concepts of *characteristic* and *definition* vs. *description.*

As pointed out by Sager, "in the process of concept formation we group the data of our perception and experience according to common elements which are called characteristics." (1990, p. 23). The notion of characteristic – an "abstraction of a property […] used for describing concepts" (ISO/FDIS 1087, 2019 (E), p. 2) – is crucial for the task of concept description. Most of the characteristics we mention in this study are what in Terminology is perceived as *essential characteristics*[105]: characteristics that cannot be separated from the *thing* itself; otherwise, the *thing* would no longer be what it actually is (Roche, 2015, p. 139). According to ISO 704, an "essential characteristic is one of a set of characteristics that is both necessary and sufficient to determine the extension of a concept" (2009, p. 7), hence, they play an essential role in the terminological work, where concepts are the core element of study.

Characteristics are essentially what allows us to define a concept. However, depending on the analysis of a given concept, characteristics have one of two functions

[105] "*characteristic* of a *concept* that is indispensable to understand that *concept*" (ISO/FDIS 1087, 2019 (E), p. 3).

in the task of concept systematisation: they are either (i) one of the characteristics composing the *set of characteristics* that mirrors the *intension* of a given concept, or (ii) the one (or several) characteristics added to this same set of characteristics, thus resulting in a different concept. The latter is known as *differential characteristic* and play an essential part in the organisation of a concept system as pointed out by many authors from different areas of studies since Aristotle – the precursor of logical theory[106], known as the earliest formal study of logic.

4.1.2. Differential characteristics

In his well-known work on the fundamentals of logic, which he called "Analytic", Aristotle laid down the basic laws of concept, characteristics, reasoning, inference, definition, to name a few, inspired by his mentor, Plato, whose work marks the beginnings of the theory of concept and epistemology (Felber, 1984, p. 102). The expression *specification by differentia* comes from his work, as well as the idea of properties – as stated in his work "categories of interpretation": Aristotle considers differences of the genus the properties that differentiate the several species of that same genus (e.g. *biped*, is the difference from the genus *animal* that differentiates the species *man* from other species of that same genus). In 1b16-24, he claims that two genera may have the same exact difference if one of them is a sub-genus of the other (Minio-Paluello, 2016).

Notwithstanding its original philosophic perspective, the differentia specification is a recognised epistemological approach in the contemporaneous work of Terminology

[106] Despite some controversial discussions on Aristotle's works, modern logicians embrace his methodology for inferential systems, as stated by Smith: "In the last century, Aristotle's reputation as a logician has undergone two remarkable reversals. The rise of modern formal logic following the work of Frege and Russell brought with it a recognition of the many serious limitations of Aristotle's logic; today, very few would try to maintain that it is adequate as a basis for understanding science, mathematics, or even everyday reasoning. At the same time, scholars trained in modern formal techniques have come to view Aristotle with new respect, not so much for the correctness of his results as for the remarkable similarity in spirit between much of his work and modern logic. As Jonathan Lear has put it, "Aristotle shares with modern logicians a fundamental interest in metatheory": his primary goal is not to offer a practical guide to argumentation but to study the properties of inferential systems themselves." (Smith, 2020).

– in the sense of scientific discipline (see Felber, 1984; Sager, 1990; ISO 704, 2009; ISO/FDIS 1087, 2019 (E)). Although Aristotle showed little interest in words throughout his work, focusing on the entity's properties instead, in our mixed methodology (Santos & Costa, 2015) – where we combine linguistic and conceptual analysis complementing each other without overlapping, a fundamental aspect within our perspective of the *double dimension of Terminology* (Costa, Silva, Barros, & Lucas Soares, 2012; Costa R. , 2013; Roche, 2014; Costa R. , 2017) – concepts may be inferred from the morphological analysis of the terms that designate them (either mono- or *polylexical*) (Ramos, Costa, & Roche, 2019). Such methodology relies on the acknowledgement of language as the vehicle of the thought, mirroring the conceptualisation – the *preverbal level* (Lino, 1987) – where cognitive operations are performed, as argued by Felber: "Concepts are mental representations of individual objects", and they serve as the "means for mental ordering (classification) and with the aid of a linguistic symbol (term, letter, graphical symbol), for communication" (Felber, 1984, p. 115).

According to Felber, the determination of a concept designated by a given term is linked to one or more concepts denoting one characteristic or several characteristics belonging to the same type, thus representing a sort of typology of characteristics. In the case of concept specification, that determination requires the identification of the determining element: a new concept is created by the addition of at least one determining concept to its genus. It must be noted that, for this author, a "characteristic is an element of a concept which serves to describe or identify a certain quality of an individual object" (1984, p. 172).

Felber also offers different approaches depending on the level of analysis. According to this author, when positioning ourselves at the level of linguistic analysis, if we take, for instance, a polylexical term constituted by two units, the term that points to the genus is the *determined member (constituent)* while the added characteristic to the genus is signalled by a co-occurrent, which is the *determining member* within the morphological structure of the term. We have resorted to Felber's example (1984, p.

172) below to demonstrate how differential characteristics might be inferred from a linguistic analysis:

> vehicle land = land vehicle
>
> vehicle = *determined member*
>
> land = *determining member* [a characteristic belonging to a specific type of characteristics, namely to the set of characteristics comprised by land, sea, air, space, etc.]

Regarding those terms with the above-demonstrated morphological composition, i.e., "land vehicle", Felber asserts that they operate as a sort of *short definition*. However, such inference is not that simple. It is one among other mechanisms that are needed to analyse natural language definitions, as we intend to demonstrate in this study. In our opinion, the *determining member* that Felber considers to be a *short definition* corresponds to what Aristotle calls the differential characteristic. However, the two authors position themselves at different levels of analysis.

As mentioned in the previous Section (4.1.1), essential characteristics are indeed necessary to define concepts: either by intension – the whole set of necessary characteristics of a given concept (see ISO/FDIS 1087, 2019 (E)) – or by extension – the whole set of concepts sharing the intension of the superordinate concept (*ibid.*). However, what determines the place of a given concept in the concept system is one characteristic or a set of characteristics that makes it unique; this determines its position in relation to other concepts, both in a horizontal relation – where the coordinated[107] concepts can be found – and in a vertical relation – where generic or specific concepts interrelate hierarchically. In Terminology, such determining characteristic is currently

[107] According to the latest version of ISO 1087 at the time of writing, a coordinated concept is a "subordinate concept resulting from the same criterion of subdivision as another subordinate concept". (ISO/FDIS 1087, 2019 (E), p. 5).

called delimiting characteristic[108], a crucial element to formulate concept definitions to which Aristotle called *differentia* in his theory of definition.

As advanced by Smith, the issue of *differentia* is at the core of Aristotle's theory of definition:

> a definition defines an essence, only what has an essence can be defined. What has an essence, then? That is one of the central questions of Aristotle's metaphysics; once again, we must leave the details to another article. In general, however, it is not individuals but rather **species** (eidos: the word is one of those Plato uses for "Form") that have essences. A species is defined by giving its **genus** (genos) and its **differentia** (diaphora): the genus is the kind under which the species falls, and the differentia tells what characterizes the species within that genus. As an example, human might be defined as animal (the genus) having the capacity to reason (the differentia). (Smith, 2020)

Smith also points out that, for Aristotle, a definition is "an account which signifies what it is to be for something" […]. The phrase *"what it is to be"* and its variants are crucial: giving a definition is saying, of some existent thing, what it is, not simply specifying the meaning of a word (Aristotle does recognize definitions of the latter sort, but he has little interest in them)" (Smith, 2020).

In 1967, Cassidy stated that a definition is:

> laying something down (72a22). Aristotle neatly distinguishes definition from hypothesis by stating that only if being were a genus, and hence definable, (and it is not a genus) could its existence be proven by definition (90b15-18). Hypothesis and definition differ further in that the formula of the latter consists at least of a term, the differentia, which characterizes (together with its implications) the specific kind of thing an object would be [Metaphysics Z. 12 (1038a 29-31)] and the formula of the first states that a subject exists, predicating an attribute of it. (Cassidy, 1967, p. 112)

[108] "essential characteristic used for distinguishing a concept from related concepts. NOTE The delimiting characteristic support for the back may be used for distinguishing the concepts 'stool' and 'chair'." (ISO 1087-1, 2000, p. 3).

For Aristotle, a definition is, therefore, an essential predication of the thing. Essential predication is about concepts and not about words, i.e., concepts are defined, not terms. Furthermore, an essential predication is also what it is said about a subject – where the difference between individuals and universals[109] arises. What we can say about an individual are its species and genus, together with the differences, because the predicative relation among them is definitory. Finally, this predicative relation involves an ontological dependency, to the extent that species and genus only exist as far as individuals exist (see Minio-Paluello, 2016).

Reading the lines above, it is clear that Aristotle coined a number of terms that are still used in knowledge organisation, particularly regarding the metalanguage of ontologies and concept definition, viz., individuals, universals, genus, species, and predicate, just to name a few.

However, there are other elements necessary for the task of writing concept definitions that are also relevant. These elements are considered *supplementary information* (ISO 704, 2009, p. 37) for the task of writing textual definitions, although they do not play an essential role in every terminological work[110]. This supplementary linguistic information allows us to hypothesize *descriptive characteristics*.

4.1.3. Descriptive characteristics

As mentioned above, some characteristics are necessary, namely, the descriptive characteristics which some authors refer to as *accidental* characteristics or *attributes* of an object (Roche, 2009, p. 13), e.g., the red colour of an apple (before turning red, the apple was green). Wüster (1998, p. 55) and Kocourek (1985, p. 124) refer to these

[109] According to Aristotle, "Subjects may be either individual or universal, but predicates can only be universals: *Socrates is human*, *Plato is not a horse*, *horses are animals*, *humans are not horses*." (Smith, 2020).

[110] According to ISO 704: "Supplementary information plays an important role in terminology databases that contain terminology for translation and writing purposes where the emphasis is on how the terminology is used in discourse. Supplementary information plays a less important role in systematic terminology work for information and knowledge management where the emphasis is on the concept system and the relations between the concepts." (ISO 704, 2009, p. 29).

accidental characteristics, such as colour and shape, as *intrinsic*[111] *characteristics*, and provide different perspectives regarding the priority of these characteristics to describe a concept. While Wüster and Kocourek consider shape and colour (just to name a few) important characteristics to describe a concept, Sager (1990) and Roche (2007) agree on the fact that such characteristics are not fundamental to understand the concept, therefore, are referred as *inessential* by Sager*:*

> The sufficient and necessary characteristics for identifying concepts are […] called essential, in contrast to inessential ones which are observable in the individual object, e.g. the colour, material, number of legs of tables. (Sager, 1990, p. 24)

Notwithstanding, despite their inessential quality for concept comprehension regarding its place in the concept system, descriptive characteristics still play a crucial role in understanding a concept at the level of abstraction. ISO 704 (2009) highlights the relevance of supplementary information. However, this standard does not consider it as being at the same level as essential information within the task of defining a given concept in natural language. Instead, it recommends that supplementary information should be stated in a separate place in the definition's text, namely as a note. This means, it should not be included in the definition itself, but adjacent to it. Finally, and according to the mentioned standard, such supplementary information plays the role of descriptive information:

> definitions should be as concise as possible and as complex as necessary. Complex definitions shall contain only information that makes the concept unique; any additional descriptive information deemed necessary is to be included in a note. (ISO 704, 2009, p. 27)

In line with this rationale, we have used some descriptive characteristics while systematising the concepts of the domain under study. Our aim was not to define concepts, given the inessential role played by descriptive characteristics within the

[111] According to Wüster, an intrinsic characteristic can be observed by simply examining a given object and does not require more knowledge about the use or origin of the object (1998, p. 55).

theory of concept (Felber, 1984, p. 103) – in the sense of ISO standards recommendations for the terminological work (ISO 704, 2009) – but to demonstrate that, at the level of specific individuals, descriptive characteristics can be useful to represent a shift of status, more specifically, an evolution within a process, as demonstrated in Section 6.5.1 (p. 276).

The outlined reflection aims at bridging three main aspects that have inspired us in our study: (i) the classical aspects of logic; (ii) the methodology of our terminological work – in which characteristics play a fundamental role in the analysis or the elaboration of intensional definitions; and finally (iii) formal definitions, for which we resorted to Protégé and inherent *Web Ontology Language* (OWL) – a W3C Recommendation[112] – to formally describe the concepts of the domain to relate them through high-level abstract syntaxes and formal reasoning[113] in a reason-*able* ontology since concepts are coherently defined, as further demonstrated in Section 6 (p. 215).

4.2. Analysis and representation of textual definitions

In this section, we analyse four different definitions that were semi-automatically extracted from the Cork Corpus constituted by normative and technical texts. The definitions identified in this corpus were written by experts. All of them define some kind of <Cork stopper>.

We retained four definitions that define the three following concepts:

1. <Stopper>

> **Definition 1**: product obtained from natural cork and / or agglomerated cork, consisting of one or more pieces, intended to seal bottles or other

[112] https://www.w3.org/TR/owl-guide/

[113] in the sense of an automated classification, a feature obtained from a reasoner like HermiT, a plugin of Protégé (ontology editor tool).

containers and to preserve their contents. (Literal translation) Source: (Cork Corpus 5.1 - NORM)

Definition 2: Piece of cork, usually cylindrical, conical or prismatic quadrangular, sometimes with rounded or chamfered lateral edges, consisting of one or several glued elements and intended to seal the containers or contribute to their water tightness. (Literal translation) (Source: Cork Corpus 7.8 – TECH)

2. <Natural cork stopper>

   Definition 3: stopper consisting entirely of natural cork.

   Note: Natural cork stoppers that have been submitted to the sealing operation (see 6.5.5) are commonly referred to as colmated natural stoppers (Literal translation) Source: (Cork Corpus 5.5 – NORM)

3. <Colmated natural cork stopper>

   Definition 4: the colmated natural cork stopper is a stopper made of natural cork whose lenticels are filled with a mixture of glues and cork powder from the dimensional finishing processes of natural cork stoppers (literal translation) Source: (Cork Corpus 6.1 – REP)

For the concept <Stopper> we retained two definitions because they contain complementary information that we are going to use to define the concept. These two definitions are referred to in this study as Definition 1 and Definition 2.

4.2.1. Linguistic analysis of Definition 1

Below follows our analysis of two definitions we found in our corpus for the concept <Stopper>. We shall start by addressing Definition 1. We wrote down the details of this analysis below in Table 17. The analysis of Definition 2 will be addressed in a

second moment and written down in Table 18. Finally, we will merge the information gathered from both definitions and arrange it on a lexical map.

In Table 17, we systematise the linguistic analysis of the textual definition. It represents the first moment of our study, where we describe the deconstruction of the textual definition and present the linguistic analysis of the definition.

Table 17: Linguistic analysis of <stopper> definition

Concept				
<Stopper>				
Definition in context				
product obtained from natural cork and / or agglomerated cork, consisting of one or more pieces, intended to seal bottles or other containers and to preserve their contents. (Literal translation) Source: (Cork Corpus 5.1 - NORM)				
LINGUISTIC DIMENSION	**Analysis**	**Lexical marker (LM)**	**Lexical-semantic relations**	**Interpretation**
	stopper [is a] product	'is a' = Ø	**HYPERNYMY - HYPONYMY**	product [GENERIC] stopper [SPECIFIC]
	stopper [consists of] one or more pieces	*'consisting of'*	**HOLONYMY-MERONYMY**	stopper [OBJECT] one or more pieces [COMPONENTS]
	stopper [is obtained from] natural cork	*'obtained from'*	**HOLONYMY-MERONYMY**	stopper [OBJECT] natural cork [STUFF]
	stopper [is obtained from] agglomerated cork	*'obtained from'*	**HOLONYMY-MERONYMY**	stopper [OBJECT] agglomerated cork [STUFF]
	stopper [is obtained from] natural cork and agglomerated cork	*'obtained from'*	**HOLONYMY-MERONYMY**	stopper [OBJECT] natural cork and agglomerated cork [STUFF]

As we can see, the definition presented in Table 17 points at two characteristics, namely (1) the compositional structure (parts) of the <Stopper> (LM = 'consisting of'); and (2) the type of substance the <Stopper> is made of (LM = 'obtained from'). Definition 1, thus, conveys two axes of analysis, namely Parts and Substance.

We mainly focused on the identification of the lexical markers (LM) and on how they express lexical-semantic relations between terms. This analysis permits us to identify the specific relation between term A and term B, which are the core elements of a textual definition. There are two kinds of such relations: hypernymy-hyponymy and holonymy-meronymy. This systematisation will allow us to finally step into modelling the information.

Modelling specialised information conveyed by a textual definition is not a straightforward task. It involves several steps and each step may depend on the previous one.

The first information that we get from the analysis is that a <Stopper> “is a product”. In this statement, “is a” is a lexical marker that relates term A “stopper” and term B "product” giving us a clear hypernymy-hyponym relation, where “stopper” is the hyponym of the hypernym “product”:

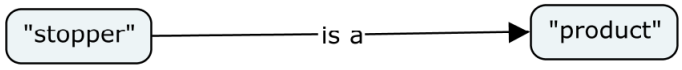


Figure 21: Representation of the lexical marker “is a” relating term A (“stopper”) to term B (“product”).

The term “stopper” is the specific term, and “product” is the generic one.

Although elided from the textual definition, the linguistic expression “is a” is inferred by the reader as existing between the term entry (the definiendum) and the definitional sentence (the definiens) – a typical feature in definition writing. The linguistic marker “is a” is considered the most common linguistic expression that denotes the lexical-semantic relation of hypernymy[114]:

[114] Hypernymy is a lexical-semantic relation of inclusion between two lexical items where the most generic sense is included in the most specific sense. According to Dubois, “L’hyponymie désigne un rapport d’inclusion appliqué non à l’objet référé, mais au signifié des unités lexicales concernés ; ainsi il y a l’inclusion du sens de *chien* dans le sens d’*animal :* on dit que *chien* est un hyponyme d’*animal.”* (2002,

a [stopper]$_{\text{hyponym / specific}}$ is a [product]$_{\text{hypernym / generic}}$

The term “stopper” designates an object that results from production, where the term “product” is a hypernym, a lexical item carrying more generic information, while “stopper” is a hyponym, a lexical item that conveys more specific meaning. What underlies the interpretation of the specific relation we identified is that “is a” expresses a predicative feature, namely that of pointing to hypernymy as aforementioned.

The second type of information we inferred from the linguistic analysis is related to the expression “product **consisting of** one or more pieces”. In this case, “consisting of” is the lexical marker, since it relates the meaning of “product” with the number of pieces it may be composed of. It offers information regarding the compositional structure of the concept <Stopper>. Compositionality is one of the sub-types of the lexical-semantic relation of holonomy-meronymy, and since “*consisting of*” identifies the relation between “stopper” – an object resulting from a production – and “one piece” or “several pieces” – the components of the “stopper” – we are facing an OBJECT-COMPONENTS sub-type.

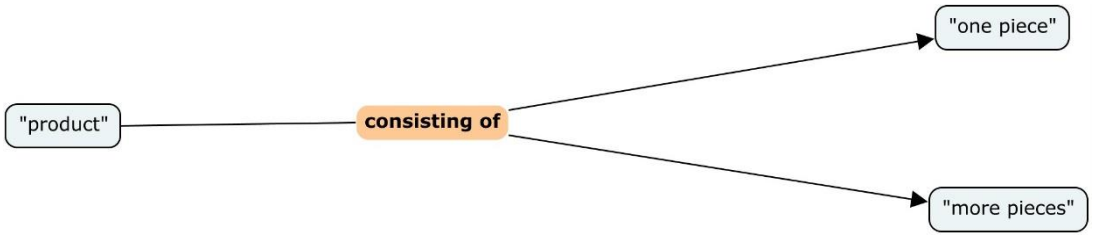


Figure 22: Representation of the lexical marker “consisting of” relating the term “product” to the information “one piece” and “more pieces”.

Figure 22 depicts the behaviour of the lexical marker “consisting of”. It relates the term “product” to “one piece” or “more pieces”. This means that the “product” –

p. 236). Cruse advocates the existence of sub-types of hypernymy: “One of the most important varieties of hyponymy (but also one of the most difficult to elucidate) is taxonymy, the relation which determines the well formedness of expressions of the form 'An X is a kind of Y', and is the vertical structuring relation of taxonomic lexical hierarchies.” (Cruse, 2002, p. 20). Nevertheless, the topic of sub-types of hypernymy will not be debated in our study.

the generic term of “stopper” – can be either composed of a single piece or several pieces.

When we particularly focused on this lexical marker relating “product” to “more pieces”, we clearly inferred that the information being pointed at is related to compositionality. This interpretation derives from the relation established between the former and the latter using the lexical marker “consisting of”: when occurring between the two linguistic forms, this lexical marker relates the meaning of “product” – the generic term of “stopper” – to the meaning of “more pieces” – a linguistic form that denotes “several parts”.

a [stopper]$_{\text{object}}$ consisting of [more pieces] $_{\text{components}}$

Considering that <Stopper> is an object that is the result of production, we assume that the term “stopper” expresses the object, while “more pieces” expresses its components (parts), thus denoting the establishment of the lexical-semantic relation of OBJECT-COMPONENTS, a sub-type of meronymy.

The third information is obtained from the interpretation of the expression “product **obtained from** natural cork”. Here, the lexical marker identified is “obtained from” and it relates the “product” to raw material, in this case, “natural cork”.

Figure 23: Representation of the lexical marker “obtained from” relating the terms “product” to “natural cork.

We can infer from this lexical marker that “product” – the generic term for “stopper” – is made of a given substance. Since the substance is the stuff of what a given object is made of, we can deduct that we are before another sub-type of the lexical-semantic relation of meronymy, namely OBJECT-STUFF.

a [stopper] $_{object}$ is obtained from [natural cork] $_{stuff}$

The OBJECT-STUFF relation was not directly inferred from the interpretation of this lexical marker. Its interpretation was complemented with the information we had previously found in the first inference drawn from the analysis of the definitional text, which can be systematised into two steps:

(i) a stopper is an object that results from production,

(ii) a stopper is made of a substance; an information conveyed by the lexical item "natural cork" – a raw material.

Similarly to what was demonstrated with the information inferred from the linguistic expression "product **obtained from** natural cork", the same applies to the term "agglomerated cork". The lexical marker "obtained from" also relates this term to "product" like we previously observed with "natural cork".

The representation of the relation established by the lexical marker "obtained from" between "product" and "agglomerated cork" is as follows:

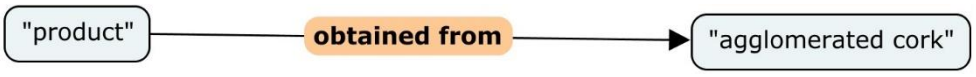


Figure 24: Representation of the lexical marker "obtained from" relating the terms "product" to "agglomerated cork".

Like "natural cork", the term "agglomerated cork" denotes a type of raw material. Here, and following the previous analysis, we can infer that a <Stopper> – a manufactured object – can be made of a different substance, other than "natural cork". In this case, the stuff of the object is "agglomerated cork", and once again we can assume that we are in the presence of a sub-type of meronymy, namely OBJECT-STUFF, which we represent as:

a [stopper] $_{object}$ is obtained from [agglomerated cork] $_{stuff}$

In this interpretation, “stopper” refers to the object while “agglomerated cork” is the stuff that the former is made of. Such assumption derives from the relation established by the lexical marker “obtained from” and the terms “product” and “agglomerated cork”, pointing here to the information concerning the substance of the object.

Further focusing on the lexical marker “obtained from”, another linguistic expression deserves our attention, namely the particle “and/or” occurring between the terms “natural cork” and “agglomerated cork”.

The co-occurrence of the linguistic expression “and/or” with the terms “natural cork” and “agglomerated cork” simultaneously expresses grammatical conjunction and disjunction of these two terms. This means that these two terms can be related to “product” either in conjunction (“natural cork **and** agglomerated cork”) or in isolation (“natural cork” **or** “agglomerated cork”).

The analysis of the relation established by the lexical marker “obtained from” and the terms “natural cork” and “agglomerated cork”, in addition to the presence of the particle “and”/”or”, led us to infer that this lexical marker relates “product” not to two terms, but three terms, given the presence of the mentioned particle. From the three relations established by the lexical marker between (1) “product” and “natural cork”, (2) “product” and “agglomerated cork”, and finally (3) “product” and “natural cork and agglomerated cork”, we can deduct that a <Stopper> is a product that can be made of three types of substances.

We decided to systematise the three types of substances pointed at by the lexical marker “obtained from”, as observed so far:

1. “product” **is obtained from** “natural cork”
2. “product” **is obtained from** “agglomerated cork”
3. “product” **is obtained from** “natural cork and agglomerated cork”

The latter we inferred from the analysis of the lexical marker “obtained from” as shown in the third line. The information **“natural cork and agglomerated cork”** is, therefore, the third type of substance that a <Stopper> can be made of, information shown by the lexical marker “obtained from”. As mentioned before, the substance is the stuff that a given object is made of; therefore, we can assume that the sub-type of meronymy is, here again, OBJECT-STUFF, as displayed below:

a [stopper] $_{\text{object}}$ is obtained from [natural cork and agglomerated cork] $_{\text{stuff}}$

We can conclude that “stopper” is an object that can be made of “natural cork and agglomerated cork”, and the latter is the stuff in the lexical-semantic relation OBJECT-STUFF.

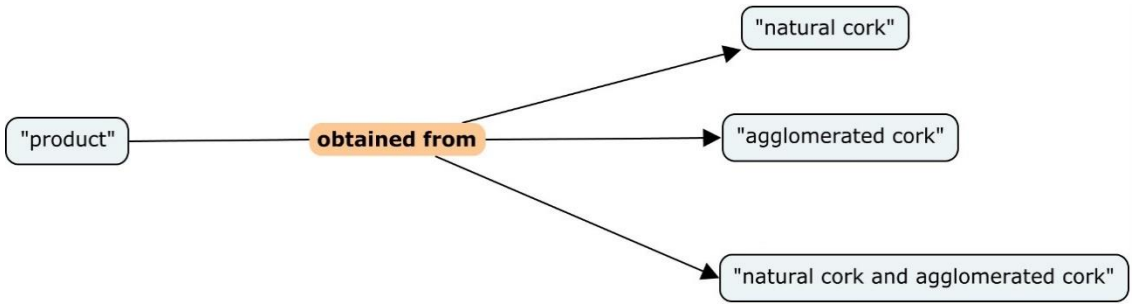
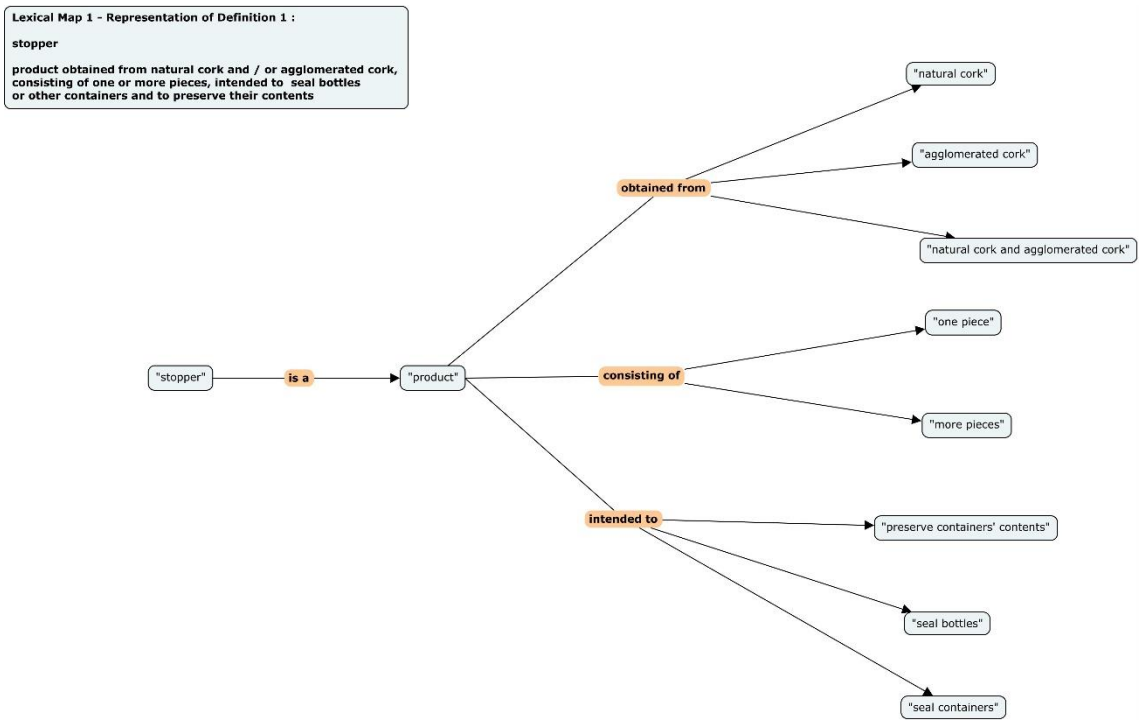


Figure 25: Representation of the lexical marker “obtained from” entertaining the same sub-type of meronymy, namely OBJECT-STUFF between “product” and “natural cork”, “agglomerated cork” and “natural cork and agglomerated cork”.

The analysis of the lexical marker “*obtained from”* was modelled on a map with three stems in order to represent the three types of information related to “product” by this lexical marker, as shown above in Figure 25. As observed, despite the presence of two terms denoting two types of substance, there is a third type of information related to the term “product” conveyed by the lexical marker “obtained from”.

The following lexical map shows all the above representations assembled in one map.



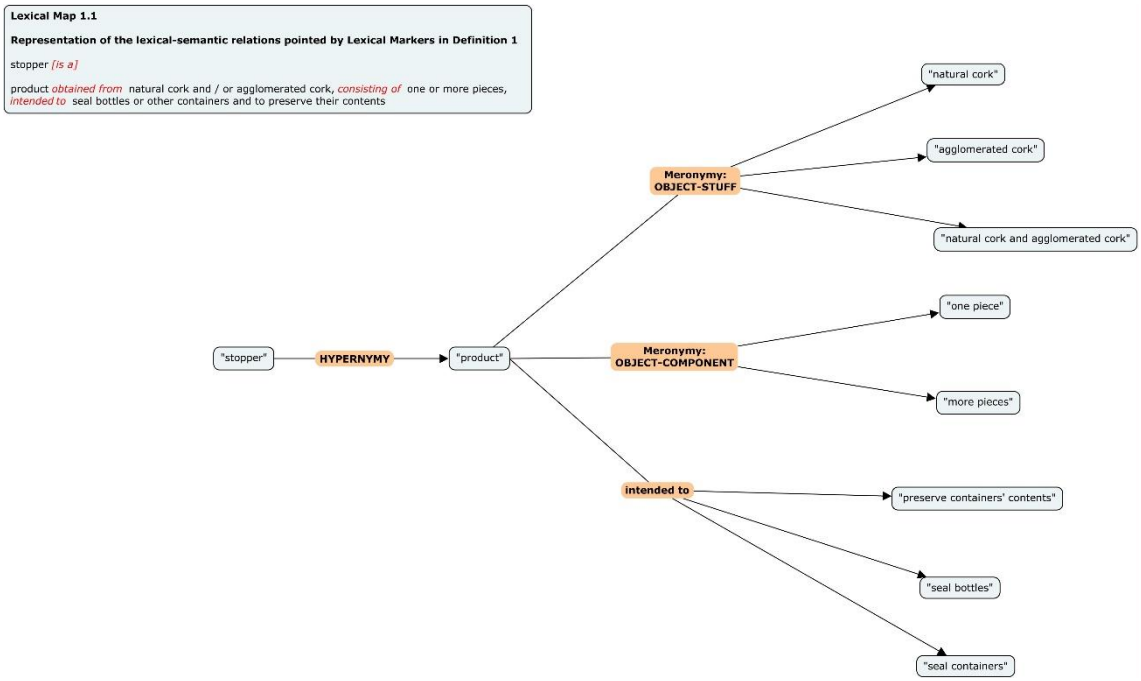
Lexical Map 1 – Representation of the interpretation of Definition 1

Lexical Map 1 corresponds to the representation of the interpretation of Definition 1 in an environment where we have modelled the lexical information obtained via the analysis of the definition, where we followed the syntagmatic order of the lexical items in the definition.

Lexical items, such as “stopper”, “product” and “natural cork”, are inserted into nodes and written with quotation marks. The arcs, where linguistic expressions are highlighted in orange, represent what we designate as lexical markers and play the role of connecting nodes. With this representation, we can observe the predicative feature of lexical markers (LM): that of connecting lexical items and specialised lexical items (terms) in a particular and recurrent[115] morphosyntactic structure that commonly underlies specialised knowledge information. The relevance of these linguistic markers

[115] Further observations in other definitions analysed in this study will demonstrate the recurrence of certain linguistic expressions.

in the terminological work is mostly regarding their capacity of pointing at lexical-semantic relations, a feature that helps the terminologist model specialised information.



Lexical Map 1.1 - Representation of the lexical-semantic relations pointed at by Lexical Markers in Definition 1

Following the representation of the interpretation of Definition 1 in Lexical Map 1, a second map was built, where lexical markers, illustrated as arcs on Lexical Map 1, are replaced by lexical-semantic relations.

Lexical Map 1.1 is a representation of the lexical-semantic relations pointed at by the lexical markers found in Definition 1. In effect, we simply replaced the lexical markers represented in Lexical Map 1. One of the interesting aspects of this kind of representation is that one can swiftly visualise the linguistic analysis demonstrated up to a specific moment. Moreover, the possibility of visualising systematised data allows us to confirm the insights of the ongoing analysis.

It is important to remark that one of the lexical markers systematised on Lexical Map 1.1, namely the linguistic expression “intended to” is not replaced by any lexical-semantic relation. Such fact is a consequence of the inexistence of a classification for

the associative[116] concept relation in Semantics, despite its verbalisation in natural language. Regardless of this inexistence of a classification, we maintained the linguistic expression on Lexical Map 1.1 given its important predicative feature in co-text with the term "stopper", i.e., it points to the function of this object.

4.2.2. Linguistic analysis of Definition 2

Definition 2 defines the same concept as Definition 1, namely <Stopper>.

Identically to Definition 1, we have recorded in Table 18, the observations of the analysis. The methodology and underlying goals are identically followed, as previously described: first, we focused on the identification of lexical markers through the analysis of the behaviour of terms on the syntagmatic axis to identify the lexical-semantic relations expressed between them.

Table 18 below represents the first moment of the analysis of Definition 2, where we systematise the deconstruction of the textual definition and present the lexical markers we have identified.

Table 18: Linguistic analysis of Definition 2: the second definition of <stopper>

Concept				
<Stopper>				
Definition in context				
piece of cork, usually cylindrical, conical or prismatic quadrangular, sometimes with rounded or chamfered lateral edges, consisting of one or several glued elements and intended to seal the containers or contribute to their water tightness (Literal translation) Source: (Cork Corpus 7.8 – TECH)				
LINGUISTIC DIMENSION	**Analysis**	**Lexical marker (LM)**	**Lexical-semantic relations**	**Interpretation**
	stopper [is a] piece of cork	'is a' = Ø	**HYPERNYMY - HYPONYMY**	piece of cork [GENERIC] stopper [SPECIFIC]

[116] According to (ISO/FDIS 1087, 2019 (E)), the associative relation is a pragmatic one, and it is considered a non-hierarchical concept relation.

	stopper [is made of] cork	'is made' = Ø 'of'	**HOLONYMY-MERONYMY**	stopper [OBJECT] cork [STUFF]
	stopper [is] cylindrical	'usually'	**HYPERNYMY - HYPONYMY**	stopper [GENERIC] cylindrical stopper [SPECIFIC]
	stopper [is] frustoconical	'usually'	**HYPERNYMY - HYPONYMY**	stopper [GENERIC] conical stopper [SPECIFIC]
	stopper [is] prismatic quadrangular	'usually'	**HYPERNYMY - HYPONYMY**	stopper [GENERIC] prismatic quadrangular stopper [SPECIFIC]
	stopper [with] rounded lateral edges	'sometimes with'	**HYPERNYMY - HYPONYMY**	stopper [GENERIC] stopper with rounded edges [SPECIFIC]
	stopper [with] chamfered lateral edges	'sometimes with'	**HYPERNYMY - HYPONYMY**	stopper [GENERIC] stopper with chamfered edges [SPECIFIC]
	stopper [consists of] one or several elements	'consisting of'	**HOLONYMY-MERONYMY**	stopper [OBJECT] one or more pieces [COMPONENTS]

As shown in Table 18, Definition 2 points at four characteristics: two identical characteristics already pointed at by Definition 1, namely, (1) the substance of which the <Stopper> is made of, however, here differently expressed (LM = 'piece of'). And (2) the compositional structure of the <Stopper> expressed by the same linguistic expression as identified in Definition 1 (LM = 'consisting of'). The two novel characteristics pointed at by Definition 2 are (3) the type of shape that a <Stopper> may be manufactured (LM = 'usually') and (4) the type of shape that a specific part of the <Stopper> may present, as a result from an operation (LM = 'sometimes with').

We will now systematise the analysis of the lexical markers and the lexical-semantic relations they express. The analysis will only focus on the two novel characteristics pointed at by Definition 2, namely *shape* and *operation*. However, a short

note must be introduced regarding the two first lexical markers recorded in Table 18, namely “is a” and “is made of”.

The first two lexical markers stated above are observed in the statement <Stopper> is a “piece **of** cork”. Here again, the lexical-semantic relation pointed at by the elided LM “is a” is hypernymy-hyponymy in the same way we observed in Definition 1: “stopper” is the specific term, and “piece” is the generic one. The underlining point here is that we have not considered “piece of cork” as the generic term. The reason for this decision ties with the analysis of this lexical item: we observed that the preposition “of” relates the term “piece” and “cork” in the same way as lexical markers do, thus, expressing a lexical-semantic relation.

Figure 26: Representation of the lexical markers “is a” and “[made] of”.

As we can see in Figure 26 above, not only the lexical marker “is a” is elided on the textual definition, but also part of the lexical marker “made of”.

In this case, the lexical marker “[made] of” relates the term “piece” – the generic term of “stopper” – with the term “cork” – a raw material. Since the latter is a type of substance that an object can be made of, we assume that the lexical-semantic relation established between these two terms is meronomy, sub-type OBJECT-STUFF.

a [piece] $_{\text{object}}$ [made] of [cork] $_{\text{stuff}}$

Finally, and considering that “stopper” is the hyponym of “piece”, we can then merge the two lexical-semantic relations pointed at by the lexical markers “is a” and “made of”, as shown in Figure 26 above, and represent the interpretation as follows:

a [stopper] $_{\text{object}}$ is [made] of [cork] $_{\text{stuff}}$

The above analysis attempts to demonstrate how lexical markers can be expressed differently or (partially) elided. In this case, we got two kinds of information

that were not explicit in the textual definition but were still inferred through the analysis of the relations established between terms using the LM.

Another information was obtained from the analysis of the text “piece of cork **usually** cylindrical”. In this statement, the lexical marker “usually” relates the term A “piece” – the generic term of “stopper” – with term B “cylindrical” in a hypernymy-hyponym relation, where “piece” is the hypernym, and “cylindrical [piece]”, the hyponym. Here we can observe another lexical marker pointing at hypernymy-hyponymy, different from the ones we have seen so far.

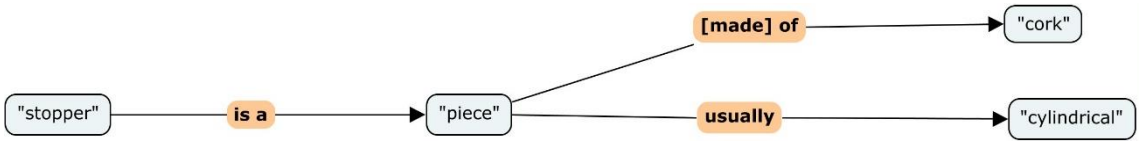


Figure 27: Representation of the lexical marker (LM) “usually”, in addition to the previous LM “is a” and “[made] of”.

As represented in Figure 27, there is a specification of the meaning of the term “piece”. This specification is an outcome of the relation established by the LM “usually” between “piece” and “cylindrical”. Since “piece of cork” is the generic term of “stopper”, we can interpret that (1) a “stopper” is a “cylindrical” piece of cork and (2) a “cylindrical piece of cork” means a “cylindrical stopper”, where the latter is a specification of “stopper”. That is, the meaning of the term “stopper” becomes more specific given the presence of the adjacent term “cylindrical” in the co-text with the lexical marker, as represented below:

a [stopper] $_{\text{Hypernym/ generic}}$ is usually [cylindrical] $_{\text{Hyponym/specific}}$

From the specification of meaning demonstrated above, we deduce that we are in the presence of hypernymy-hyponymy, in which the inferred “cylindrical stopper” falls in the category of hyponym, a more [SPECIFIC] term. On the other hand, “stopper”, a more [GENERIC] term, is the hypernym of this lexical-semantic relation expressed by the LM “usually”.

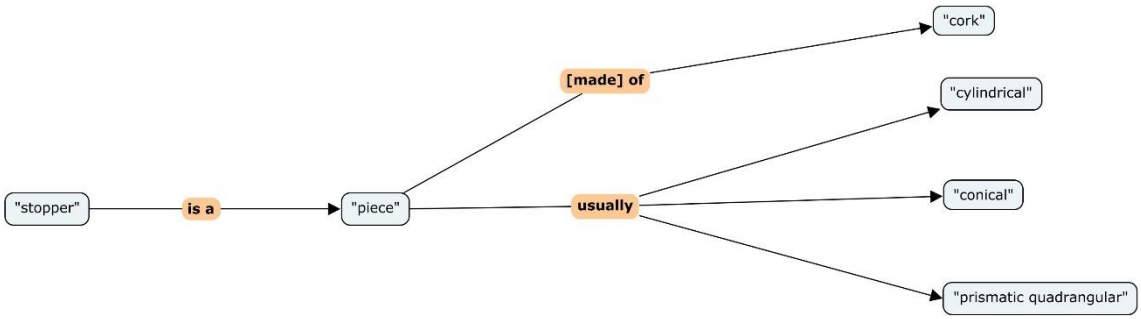


Figure 28: Representation of the lexical marker "usually", relating the term "piece" with 3 terms: "cylindrical", "conical" and "prismatic quadrangular".

The same interpretation applies to the terms "prismatic quadrangular" and "frustoconical". These two terms, and identically to "cylindrical", denote a type of shape. As depicted above in Figure 28, we systematised the information under the form of three stems, which correspond to the lexical-semantic relations established between "piece [made] of cork" and the three types of shape that a <Stopper> might present.

The next information we have inferred was obtained from the interpretation of the statement "piece of cork, **sometimes with** rounded lateral edges". The lexical marker is "sometimes with" and relates the term "piece of cork" – the generic term of "stopper" – and the term "rounded lateral edges".

In this case, we get the information regarding the shape of a specific part of the <Stopper> through the relation established by the LM "sometimes with" between the term "piece" and the linguistic form "rounded lateral edges". The LM relates the information conveyed by this linguistic structure and the term "piece" in such a way that the latter gets additional meaning. That is, a specification of "piece" is apprehended through the supplementary meaning conveyed by "rounded lateral edges", from which we assume that the lexical-semantic relation observed here is hypernymy-hyponymy:

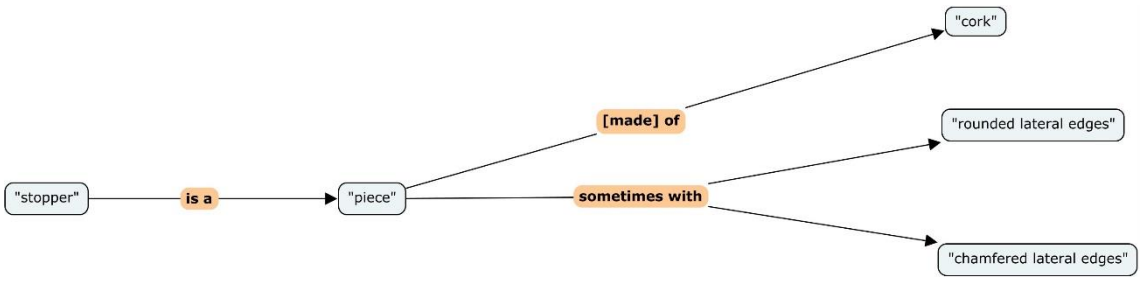


Figure 29: Representation of the lexical marker "sometimes with", relating the term "piece" with the terms "rounded lateral edges" and "chamfered lateral edges".

The same analysis applies to the term "chamfered lateral edges", as represented above, in Figure 29.

In this line of thought, the term "piece [made] of cork" – the generic term of "stopper" – is the hypernym, and "piece [made] of cork with chamfered lateral edges" is the hyponym. We can represent this interpretation, as follows:

(1) a [piece of cork] $_{\text{hypernym/generic}}$ is sometimes a [piece of cork with rounded lateral edges] $_{\text{hyponym/specific}}$

Moreover, building from the knowledge that "stopper" is the hyponym of "piece of cork", we can reformulate the previous representation as:

(2) a [stopper]$_{\text{hypernym/generic}}$ is sometimes a [stopper with rounded edges] $_{\text{hyponym/specific}}$

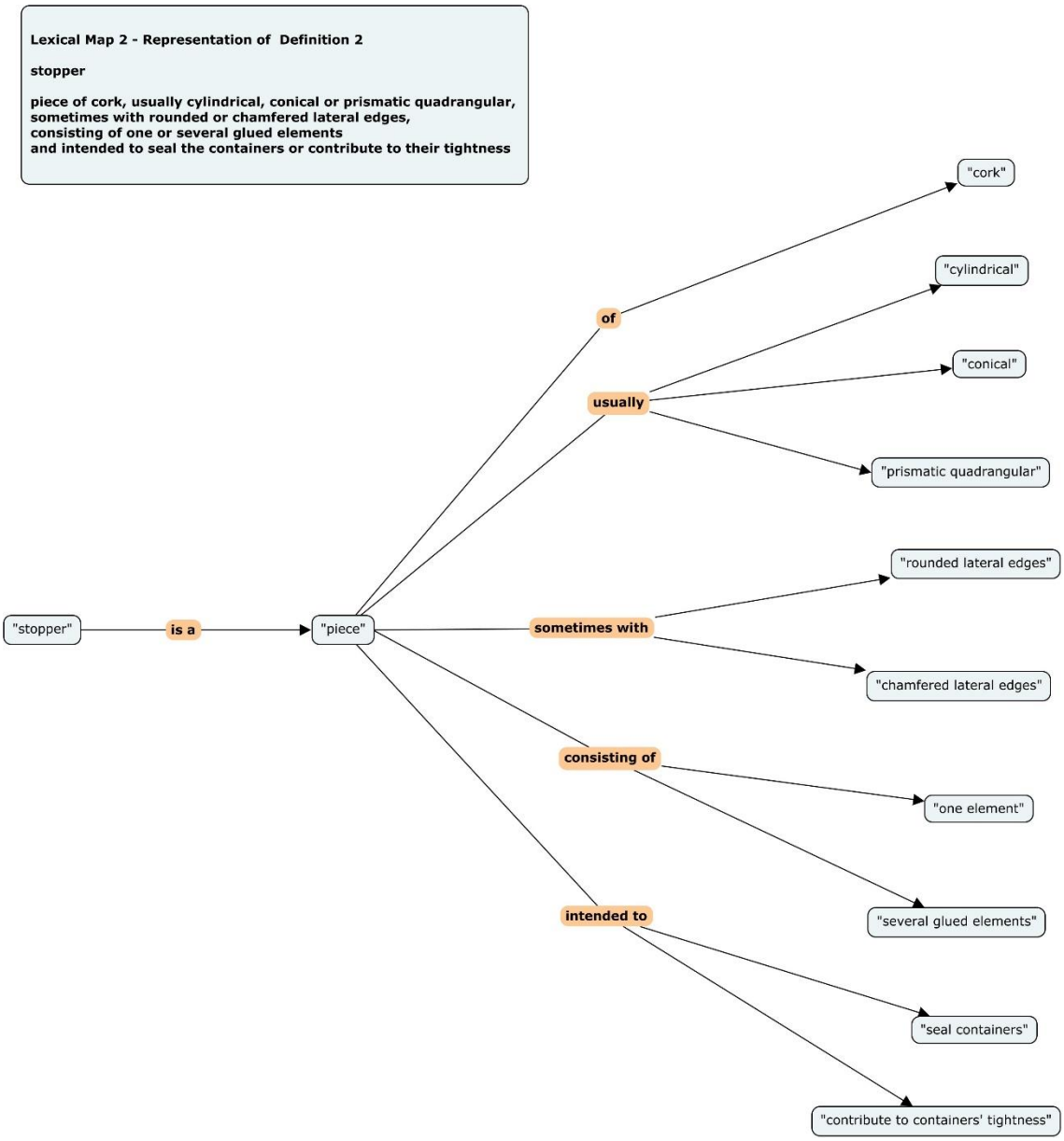
As we can see on the second representation, on which we replaced the generic term "piece of cork" by its specific term "stopper", the meaning has not changed, and the lexical-semantic relation of hypernymy-hyponymy is coherently present. Hence, the lexical marker "sometimes with" points at the lexical-semantic of hypernymy-hyponymy, where "stopper" is the [GENERIC] term and "stopper with rounded edges" = [SPECIFIC] the specific one.

In sum, we can interpret from the text of Definition 2 that a <Stopper> may have a rounded or chamfered shape on a specific part, more specifically on the “lateral edges”, in addition to the “cylindrical” main shape.

At this point, we should underline that the term “rounded lateral edges” points at a result of a type of operation that cork stoppers might be submitted to during cork stopper manufacturing. This topic is a piece of information that we acquired during the task of collecting texts, within the main task of the corpus creation. While reading and assembling specialised texts, we gained some knowledge on the domain, which we will choose to call a degree of familiarisation. It is this familiarisation that activates our awareness for non-explicit information within the textual definition.

Hence, Definition 2 does not solely focus on the substance, part and function of the object, as Definition 1 does. Instead, it also introduces the notions of shape and operation, where the former is closely related to the latter. Considering the three notions conveyed by Definition 1, we have finally assembled a set of 5 axes of analysis, namely substance, part, function, shape and operation. These five axes of analysis are at the core of the domain-ontology building, hence the relevance of keeping two textual definitions defining the same concept.

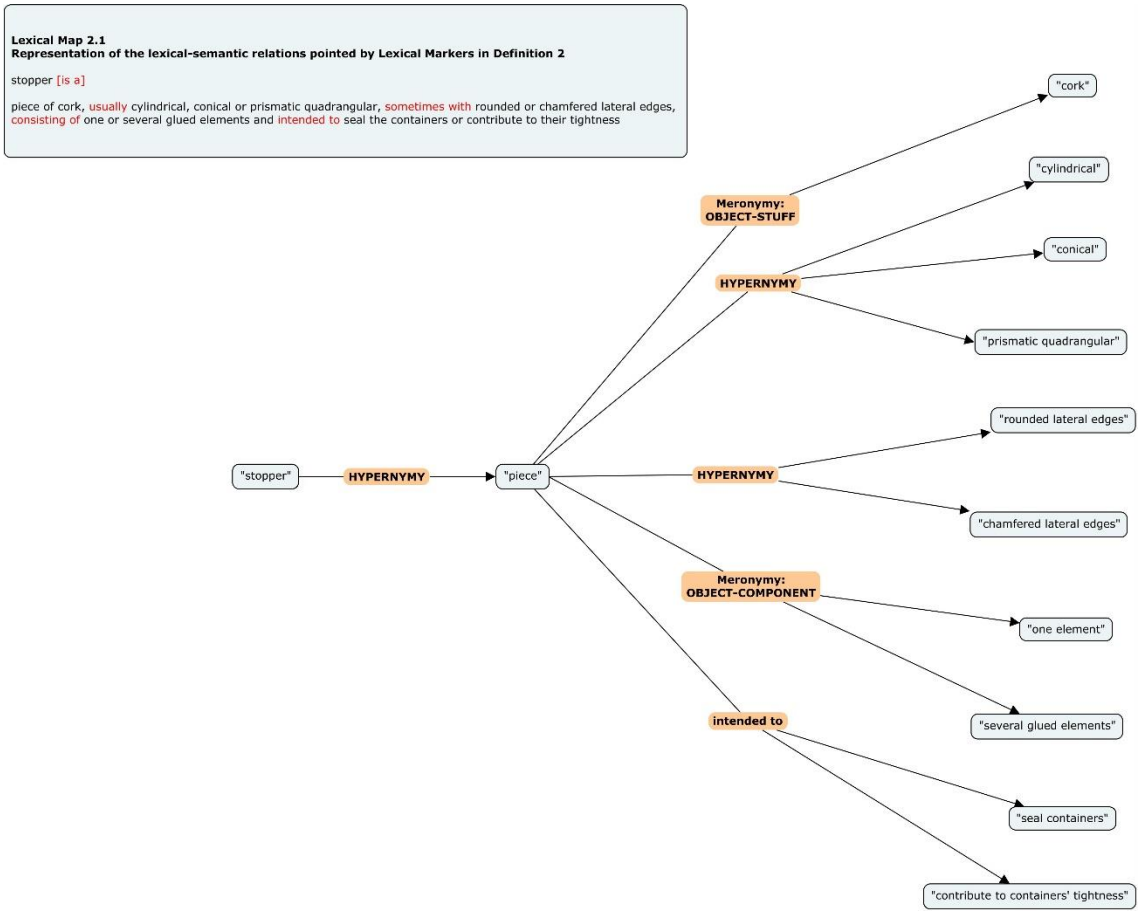
Following the methodology stated for Definition 1, we elaborated a lexical map to represent the behaviour of lexical items in the syntagmatic axis. The following lexical map shows all the above representations assembled in one single map, along with the ones where lexical markers are identical to those we had already identified in Definition 1.



Lexical Map 1 – Representation of the interpretation of the text of Definition 2.

Lexical Map 2 corresponds to the representation of the interpretation of Definition 2. Once again, we have modelled the lexical information obtained from the analysis of the definition in a map-like environment and followed the syntagmatic order of the lexical items in the textual definition. Lexical items are inserted in nodes and written with quotation marks, while linguistic markers are their connecting arcs, highlighted in orange.

After the representation of the interpretation of Definition 2 in Lexical Map 2, a second map was built, in which lexical markers, illustrated as arcs on Lexical Map 2.1, are replaced by lexical-semantic relations.



Lexical Map 2.1: Representation of the lexical-semantic relations pointed at by Lexical Markers in Definition 2.

Lexical Map 2.1 is the representation of the lexical-semantic relations pointed at by the lexical markers found in Definition 2. Once again, the linguistic expression "intended to" is not replaced by any lexical-semantic relationship, due to the inexistence of a classification for this type of relation, as mentioned before in the analysis of Definition 1 (Section 4.2.1, p. 132).

Comparing Lexical Map 2 and Lexical Map 2.1, we can see that lexical markers may be expressed by different linguistic expressions although pointing to the same lexical-semantic relation, such as the ones shown in Table 19 (below):

Table 19: Three different linguistic expressions sharing the role of lexical markers pointing at hypernymy-hyponymy

Linguistic expression	**Lexical-semantic relation**
is a (elided)	hypernymy
usually	hypernymy
sometimes with	hypernymy

However, as we will demonstrate in the conceptual analysis of Definition 2, their corresponding conceptual relations fall in a different scope of hierarchical dependency, given the pragmatic nature of the relationship. We have observed that some of the hypernymy-hyponymy relations, like the ones written above in Table 19, instead of corresponding to subsumption – a hierarchical conceptual relation that one might think of straightforwardly – do not correspond to such concept relation. Instead, they correspond to associative concept relations. This means that concepts interrelate not in a hierarchical dependency in the concept system, but in a non-hierarchical associative[117] dependency, i.e., depending on the pragmatic aspect involved (e.g., based on the cause-effect criterion[118]).

The systematisation of these observations is recorded in Table 27, Section 5.5. (p. 210).

4.2.3. Linguistic analysis of Definition 3

[117] According to ISO/FDIS 1087, an associative concept relation is a pragmatic relation and is considered a non-hierarchical concept relation (2019 (E), p. 6).

[118] This criterion is at the basis of a *sequential relation*, a sub-type of associative relation. As stated by ISO CT 37, this relation is an "associative *relation* by which *concepts* can be ordered by a relevant ordering criterion. Note 1 to entry: Sequential relations are usually based on spatial *relations*, temporal *relations* or causal relations." (ISO/FDIS 1087, 2019 (E), p. 6).

We will now address the analysis of Definition 3. This textual definition defines the concept <Natural cork stopper>.

Following our methodology as mentioned earlier, we start by analysing the behaviour of lexical items on the syntagmatic axis to identify lexical-semantic relations between terms, which are expressed by lexical markers.

Once again, the textual definition was deconstructed in order to facilitate the identification of the lexical markers and corresponding interpretation. The observations of the linguistic analysis are systematised below in Table 20.

Table 20: Linguistic analysis of Definition 3: a textual definition of the concept <Natural cork stopper>

Concept				
<Natural cork stopper>				
Definition in context				
stopper consisting entirely of natural cork Note: Natural cork stoppers that have been submitted to the sealing operation (see 6.5.5) are commonly referred to as colmated natural stoppers (Literal translation) Source: (Cork Corpus 5.5 – NORM)				
LINGUISTIC DIMENSION	**Analysis**	**Lexical marker (LM)**	**Lexical-semantic relations**	**Interpretation**
	natural cork stopper [is a] stopper	'is a' = Ø	**HYPERNYMY - HYPONYMY**	stopper [GENERIC] natural cork stopper [SPECIFIC]
	natural cork stopper [consists entirely of] natural cork	'consisting entirely of'	**HOLONYMY-MERONYMY**	natural cork stopper [OBJECT] natural cork [STUFF]
	natural cork stopper [is submitted to] the sealing operation	'submitted to'	**HOLONYMY-MERONYMY**	sealing operation [ACTIVITY] ? = [FEATURE]
	colmated natural stopper [is a] natural cork stopper	'commonly referred to as' same as = 'is a'	**HYPERNYMY - HYPONYMY**	natural cork stopper [GENERIC] colmated natural stopper [SPECIFIC]
	colmated natural stopper [results from] the sealing operation	results from = opposite of 'submitted to'	**HOLONYMY-MERONYMY**	sealing operation [ACTIVITY] colmated = [FEATURE]

As shown in Table 20, Definition 3 is written in two sentences: (1) the main sentence and (2) a footnote. The second sentence conveys essential information to understand what a <Natural cork stopper> is when submitted to a specific operation. This definition points at the substance (LM = “consisting entirely of”) and the operation (LM = “submitted to”), as systematised above. The latter characteristic, contrarily to Definition 2, is made explicit in the textual definition.

We will not address the lexical marker “consisting entirely of” in much detail since it identically expresses the lexical-semantic relation of meronymy, sub-type [OBJECT-STUFF], as previously observed with the LM “consisting of”, in Definitions 1 and 2.

However, we will still dedicate a few lines to this lexical marker to support subsequent observations.

Figure 30: Representation of the lexical markers “is a” and “consisting entirely of”.

As observed on the above representation, the LM “consisting entirely of” relates the terms “natural cork” and “stopper”. The former’s meaning points at the substance of the object, while the latter, refers to the object. We have no doubts that the lexical-semantic relation established between these two terms is meronymy, sub-type [OBJECT-STUFF], and can be represented as follows:

[stopper]$_{\text{OBJECT}}$ consisting entirely of [natural cork] $_{\text{STUFF}}$

From this representation, we can evolve into another interpretation. If we consider that “stopper” is the generic term of “natural cork stopper” – a piece of information we got from the elided “is a”, in the textual definition – we can reformulate the information in the following representation:

[natural cork stopper] $_{\text{OBJECT}}$ consisting entirely of [natural cork] $_{\text{STUFF}}$

This is the information we got from the first sentence of the textual definition.

On the second sentence – inserted as a footnote in the textual definition – another information was obtained from the analysis of the statement "natural cork stoppers that have been **submitted to** sealing operation". The lexical marker under focus is "submitted to" and it relates the term "natural cork stopper" to the term "sealing operation" as represented below:

Figure 31: Representation of the lexical marker "submitted to".

As we can see, the term "sealing operation" – which indicates an operation/activity – is related by the LM "submitted to" to the term "natural cork stopper" – which we already know to be an object. The interpretation of their meanings allows us to infer that the lexical-semantic relation established by the LM is meronymy, sub-type [ACTIVITY-FEATURE]:

[natural cork stopper] $_{\text{OBJECT}}$ is submitted to [sealing operation] $_{\text{ACTIVITY}}$

The representation above is the first moment of the inferring process, where "sealing operation" is the [ACTIVITY]. The term that points at the meaning of [FEATURE] – in the sense of resulting feature – was identified later. That is, we had to take into consideration an intermediate lexical-semantic relation observed in the continuum of the sentence, so we could reach the term that points at the meaning of [FEATURE], within the sub-type relation of meronymy, [ACTIVITY-FEATURE]. In order to achieve this goal, it was first necessary to identify the terms being related by the LM "are referred to as" and then interpret the meaning of the identified terms. The LM "are referred to as" relates the term "colmated natural stopper" with the term "natural cork stopper", as represented below:

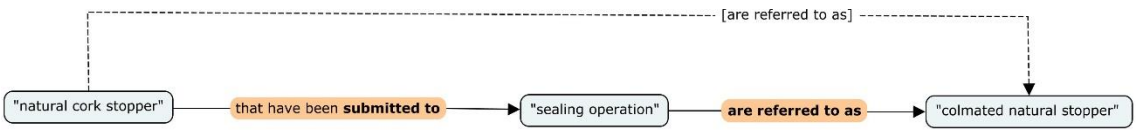


Figure 32: Representation of the lexical marker "are referred to as".

Once we identified the related terms, we deduced that the LM “are referred to as” expresses the lexical-semantic relation of hypernymy-hyponymy since the meaning of the term “natural cork stopper” is broader than the meaning of the term “colmated natural stopper”. This interpretation can be represented by:

[colmated natural stopper] $_{\text{hyponymy/specific}}$ is a [natural cork stopper] $_{\text{hypernymy/generic}}$

It was only after this observation that we finally reached the term that conveys the meaning of a result [FEATURE] within the lexical-semantic relation of meronymy, sub-type [ACTIVITY-FEATURE] – the first moment of the analysis in which we address the LM “submitted to”. That term is “colmated natural stopper” and points at a result from an [ACTIVITY], which in turn, is pointed at by the term “sealing operation”. We can now represent this interpretation, as follows:

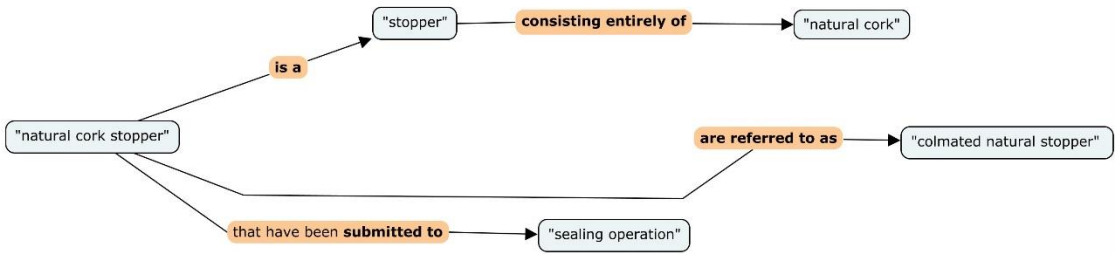
[colmated cork stopper] $_{\text{FEATURE}}$ results from [sealing operation] $_{\text{ACTIVITY}}$

Similarly to how we addressed Definitions 1 and 2, we have elaborated a lexical map in which all the above representations of the analysis of the textual definition are merged into a single map.

Lexical Map 3 - Representation of Definition 3 :

natural Cork Stopper

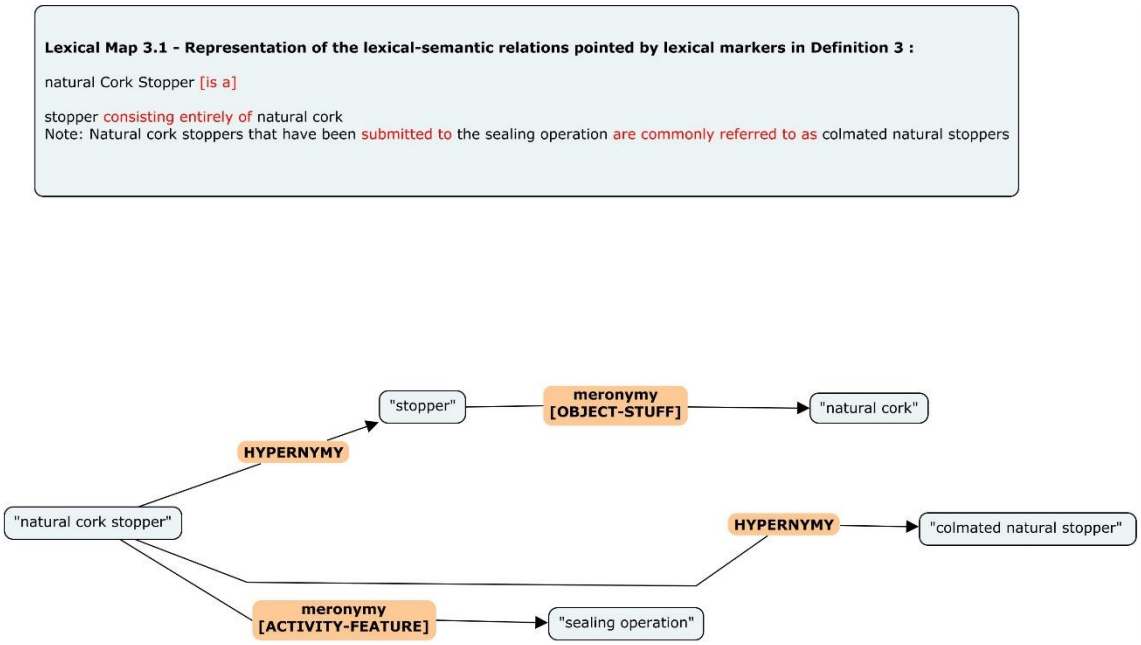
stopper consisting entirely of natural cork
Note: Natural cork stoppers that have been submitted to the sealing operation are commonly referred to as colmated natural stoppers



Lexical Map 3 - Representation of the interpretation of Definition 3: “natural cork stopper”

Lexical Map 3 unfolds into two parts in order to replicate the structure of the text in Definition 3: there is the main statement and a secondary statement, which was inserted as a footnote. In our view, this textual definition creates problems for knowledge organisation given the two objects being defined. Each definition should define a single concept; however, each of the two statements found in Definition 3 describes a different type of <Natural cork stopper>. We will further discuss this issue during the corresponding conceptual analysis and elaboration of the conceptual maps (Section 5.3., p. 195).

After the representation of the interpretation of the textual Definition 3 in Lexical Map 3, a second map was built, in which lexical markers, illustrated as arcs on Lexical Map 3.1, are replaced by lexical-semantic relations.



Lexical Map 3.1 - Representation of the lexical-semantic relations pointed at by Lexical Markers in Definition 3

As we can see on Lexical Map 3.1, identical lexical-semantic relations can be expressed by different linguistic expressions, e.g., hypernymy-hyponymy is expressed either by LM = “is a” or LM = “is referred as”. The same occurs regarding the sub-types of meronymy demonstrated so far. These observations are recorded below in Table 21:

Table 21: Lexical markers pointing at meronymy, extracted from Def. 1, Def. 2 and Def. 3

Definition	**Linguistic expression**	**Lexical-semantic relation**
definition 1	obtained from	meronymy [OBJECT-STUFF]
definition 3	submitted to	meronymy [ACTIVITY-FEATURE]
definition 1, 2 and 3	consisting entirely of	meronymy [OBJECT-STUFF]

The sub-types of meronymy written in Table 21 are no different from the ones observed in the analysis of Definition 2 and 3. The novelty found on Definition 3 is the lexical-semantic relation expressed by the lexical marker "submitted to". Although not from the analysis of the terms directly related by this LM, we were able to infer a piece of information regarding the result of an operation.

Concluding the linguistic analysis of Definition 3, we would like to highlight that text interpretation and subsequent identification of lexical-semantic relations between terms is a task that requires several steps. As demonstrated above, the lexical-semantic relation of meronymy, sub-type [ACTIVITY-FEATURE] was not straightforwardly inferred. We had to examine the behaviour of other terms in co-text with the linguistic marker under focus, for the meaning of terms is not construed through the identification of a single lexical-semantic relation established between term A and term B, nor on this specific order. Instead, the meaning of two terms can be construed in addition to the meaning of other terms co-occurring in the syntagmatic axis, as we did during the analysis of the second sentence of Definition 3: term C was not related with term B, but instead with term A. This was the case of the LM = "are referred as" and the terms "colmated natural cork" = term C, "natural cork stopper"= term A, and "sealing operation" = term B (see Figure 32, p.154).

4.2.4. Linguistic analysis of Definition 4

Definition 4 is the last textual definition which we thoroughly address to demonstrate our methodology.

Similarly to what was demonstrated with Definitions 1, 2 and 3, we first analyse the behaviour of lexical items on the syntagmatic axis of the definitional text in order to identify lexical-semantic relations between terms. The text of the definition was once again deconstructed for the identification of lexical-semantic relations pointed at by the meaning of two terms where the lexical marker plays the fundamental role of relating their meanings.

The observations of this analysis are recorded below in Table 22.

Table 22: Linguistic analysis of Definition 4: a textual definition of the concept <Colmated natural cork stopper>

Concept				
<Colmated natural cork stopper>				
Definition in context				
the colmated natural cork stopper is a stopper made of natural cork whose lenticels are filled with a mixture of glues and cork powder from the dimensional finishing processes of natural cork stoppers (literal translation) Source: (Cork Corpus 6.1 – REP)				
LINGUISTIC DIMENSION	**Analysis**	**Lexical marker (LM)**	**Lexical-semantic relations**	**Interpretation**
	colmated natural cork stopper [is a] stopper	'is a'	**HYPERNYMY - HYPONYMY**	stopper [GENERIC] colmated natural cork stopper [SPECIFIC]
	colmated natural cork stopper [is made of] natural cork	'is made of'	**MERONYMY-HOMONYMY**	colmated natural cork stopper [OBJECT] natural cork [STUFF]
	colmated natural cork stopper in which its lenticels [are filled with] cork powder	'are filled with'	**MERONYMY-HOMONYMY**	cork powder filling = [ACTIVITY] filled lenticels = [FEATURE]

	cork powder [results from] the dimensional finishing processes of natural cork stoppers	results = Ø + 'from'	**MERONYMY-HOMONYMY**	dimensional finishing process = [ACTIVITY] cork powder = [FEATURE]

This final analysis, and especially the statement that introduces information regarding what "cork powder" is and its provenance, is critical to demonstrate that there are references to "dimensional finishing processes".

Definition 4 is one of the few definitions extracted from the Cork Corpus that provided us with the notion of finishing processes. In addition to this notion, substance and operation are also pointed at by this textual definition as systematised above in Table 22.

In the following lines, we will not address the lexical markers that were already analysed in Definitions 1, 2 and 3, namely "is a" and "is made of", since the lexical-semantic relations established between the terms they relate are identical here: hypernymy-hyponym for the first LM, and meronymy, sub-type [OBJECT-STUFF], for the second.

One of the novel pieces of information found in this textual definition was introduced by the statement "whose lenticels **are filled with** a mixture of glues and cork powder", in which the lexical marker is "are filled with" and the related terms are "cork powder" and "lenticels". We will not focus on the term "mixture of glues", unless for its inclusion in the representation below.

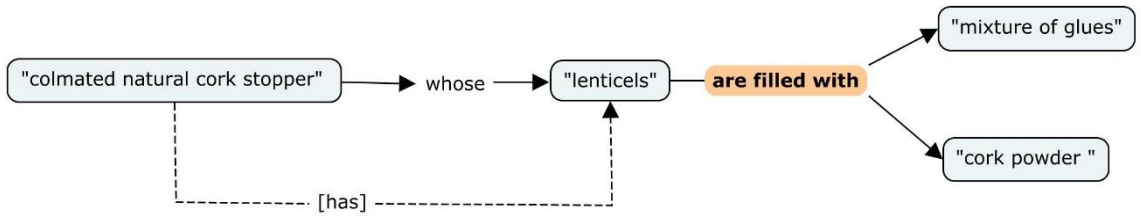


Figure 33: Representation of the lexical marker "are filled with".

As we can see in Figure 33, a twofold analysis was necessary given the possessive "whose":

(1) the analysis of the meaning pointed at by the linguistic form "whose" between "lenticels" and "colmated natural cork stopper" provides the information that the latter "has" "lenticels. From this interpretation, we could infer that

(2) "lenticels" is a term pointing at the meaning of a feature of the <Colmated natural cork stopper>.

The lexical-semantic relation inferred from the meaning pointed at by the terms "lenticels" and "cork powder" is closely related to the sense of action conveyed by the lexical marker "are filled with". We can here identify an agent (cork powder filling) and a recipient (lenticels), and finally infer an outcome of the action (filled lenticels). From this interpretation, we conclude that the term "[filled] lenticels" points at the meaning of a result [FEATURE] and the term "cork powder [filling]" points at the meaning of an operation/[ACTIVITY]. Therefore, the lexical-semantic relation observed here is meronymy, sub-type [FEATURE-ACTIVITY], as represented below:

"[filled] lenticels" FEATURE are filled with "cork powder [filling]" ACTIVITY

The prime novelty highlighted by Definition 4 is the reference to finishing processes, which is a piece of information that was inferred from the interpretation of the statement "cork powder **from the** dimensional finishing processes". The lexical marker is expressed by the linguistic expression "from the" and the terms related by this LM are "cork powder" and "dimensional finishing processes" as represented below:

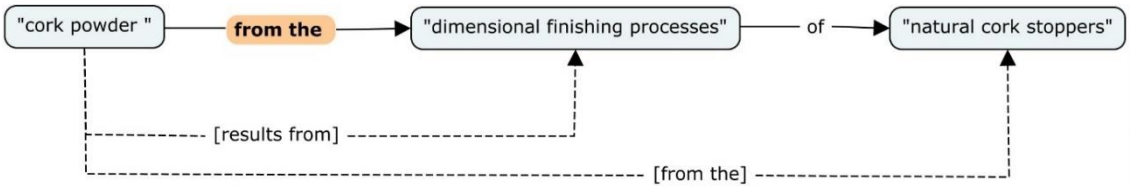


Figure 34: Representation of the lexical marker "from the".

The above representation also shows a twofold analysis. Here again, we had to examine the behaviour of another term in co-text, namely “natural cork stoppers” given the connective property established by the linguistic form “of” between “natural cork stopper” and “dimensional finishing processes”. Thus, we were able to obtain three pieces of information that are interdependent, namely:

(1) the origin of <Cork powder>;

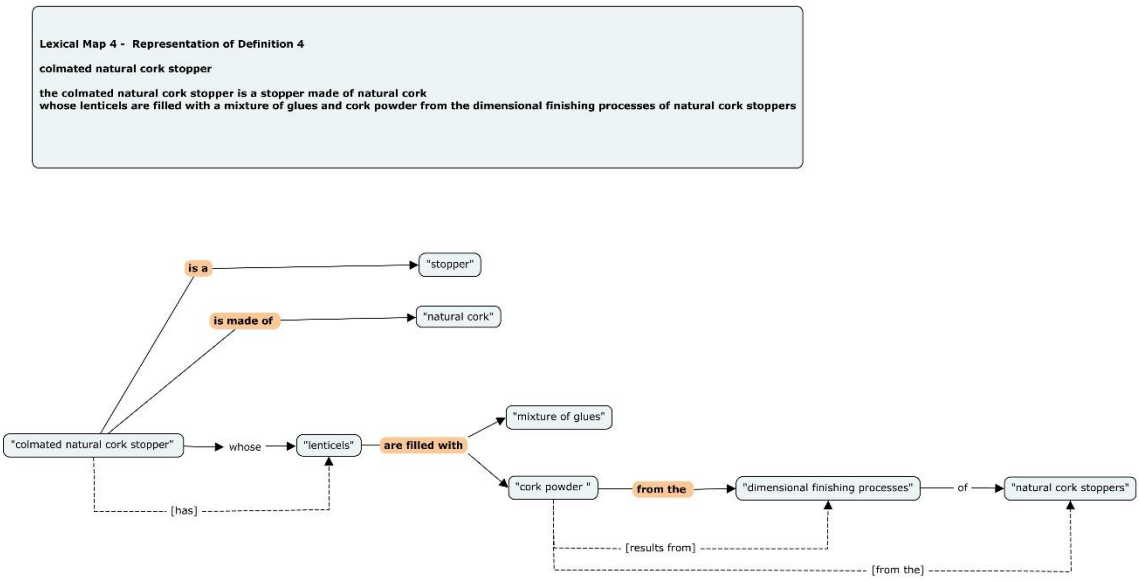
(2) the term “dimensional finishing processes” points at the meaning of an activity/operation to obtain <Cork powder>, and finally

(3) the term “cork powder” points at the meaning of a substance that results from an activity/operation.

From the outlined interpretation, we can conclude that “cork powder” is a term pointing at the meaning of a result [FEATURE] while “dimensional finishing processes” points at the meaning of an [ACTIVITY]. Therefore, a lexical-semantic relation of meronymy, sub-type [FEATURE-ACTIVITY] is in place. The representation of this interpretation is the following:

“cork powder” [FEATURE] from the “dimensional finishing processes” [ACTIVITY]

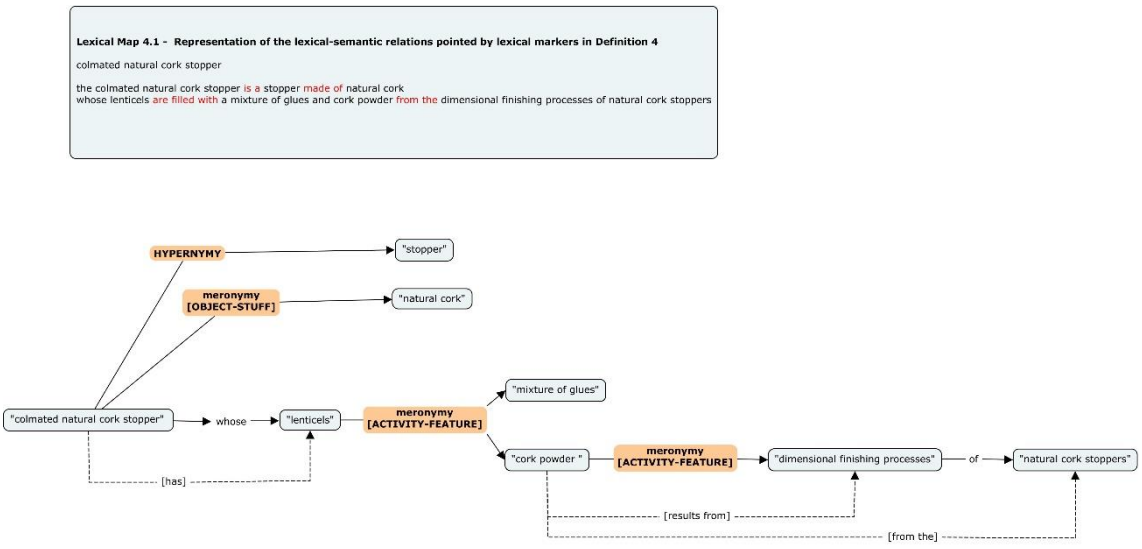
The next lexical map (4) represents the interpretation of Definition 4. We have assembled on this map all the representations that have been shown so far for the lexical makers identified in this textual definition.



Lexical Map 4 – Representation of the interpretation of Definition 4

Albeit not discussed in this analysis, the lexical markers "is a" and "is made of" are also included in Lexical Map 4. This way, we can represent the entire text of Definition 4.

Finally shown below is the lexical map where the lexical markers are replaced by the corresponding lexical-semantic relations.



Lexical Map 4.1 – Representation of the lexical-semantic relations pointed at by lexical markers in Definition 4

As mentioned before, lexical markers can be expressed by different linguistic expressions, albeit expressing the same lexical-semantic relation. The underlying argument here is that it is the meaning of the terms related by a given LM that determines the type of lexical-semantic relation established. Additionally, sometimes a further analysis of the terms' co-text must be considered, namely the linguistic forms occurring on the left- and right-hand side of the terms under focus, for the connective properties of those linguist forms co-occurring with other terms in the syntagmatic axis, allow us to construe additional meaning to the terms we are analysing. This was the case of the linguistic forms "whose" and "of" – from which we perceived the meaning of a (resulting) feature, on the other hand, pointed at by the terms "lenticels" and "cork powder", respectively (see Figure 33, p. 159).

Concluding this section, the linguistic analysis of Definition 4 did not provide us with much more information beyond what was interpreted and represented in the analysis of Definition 3 ("natural cork stopper" and "colmated natural stopper", the latter mentioned as a footnote). The relevance of Definition 4 relies on the fact that it is a textual definition of only one concept, which, in our opinion, is an aspect of significant importance. As mentioned above, each concept should be discretely treated, in as much as one clearly understands its place in a concept system. It is this clear understanding that reduces individual interpretations, on the one hand, and increments the quality of communication, on the other.

4.3. The relevance of lexical markers for modelling special knowledge information

To sum up, lexical markers are linguistic expressions that commonly point at lexical-semantic relations with a prime terminological goal: they provide us with coordinates that guide us through the task of organising knowledge information. Many authors refer to this commonality as knowledge patterns found in knowledge rich

contexts (KRC)[119] and state that some of these patterns share the same value in different fields of knowledge – i.e., they point at the same lexical-semantic relation regardless of the context[120]; whereas some others are context dependent if we think of domain-specific fields of interest (see Meyer, 2001; L'Homme, 2004; Marshamnn, 2007).

In our study, the linguistic analysis of lexical markers found in co-text with terms is paramount for modelling specific domain knowledge. It is the starting stage of the terminological work, an approach that allows us to capture information that is not obvious or that can even be suppressed. Natural language definitions written by experts do not convey all information necessary for knowledge organisation; hence, the need for an in-depth linguistic analysis of the morphosyntactic behaviour of all lexical items present in the definitional context. Only after this linguistic analysis, will we finally be able to work on the conceptual dimension, since the latter is based on the former. Therefore, this methodology implies different tasks:

(1) a linguistic analysis of definitions in natural language is carried out for the identification of specialised information, stemmed from the linguistic expressions in co-text with terms;

(2) the systematisation of the lexical-semantic relations pointed at by the lexical markers found in definitions and definitional contexts;

(3) based on the systematisation in (2), modelling the linguistic information in the form of lexical maps;

(4) based on the lexical-semantic relations systematised in (2), the identification and systematisation of the corresponding conceptual relations;

---

[119] After Meyer (2001), and followed by many authors (Condamines, 2005), (Marshamnn, 2007), (Marshamnn, L'Homme, & Surtees, N/A), (Barrière, 2004), (Halskov & Barrière, 2010), to mention only a few.

[120] A feature designated by "transportability" of knowledge patterns (see Halskov & Barrière, 2010). This topic is not addressed in our study.

(5) based on the systematisation in (4), modelling the conceptual relations in the form of conceptual maps; an operation that guides us into building an ontology, our final task.

Points (1) to (5) are the building blocks of our terminological work: from the analysis of natural language texts to the creation of a (conceptual) knowledgebase resource. Up until now, we have focused on Points (1), (2) and (3). The following lines will focus on Points (4) and (5).

Conceptual analysis

5. Conceptual analysis

Before we start describing the conceptual analysis we have made, a few introductory lines are necessary in order to justify our terminological choices. We will rely on the conceptual relations found in the conceptual analysis of Definition 1 (further demonstrated in Section 5.1) to address this matter.

The conceptual relations identified during the analysis of Definition 1 systematised below in Table 23 are based on the lexical markers we have identified in the linguistic analysis of this definition. As pointed out by Roche (2009), some conceptual relations may seem at first glance isomorphic with lexical-semantic relations – an assumption that might be identified when looking at the conceptual relations we have systematised. For instance, hypernymy vs. subsumption seems to share an identical hierarchical role in a set of two individuals/objects/abstractions holding these two types of relations – where one is always more generic than the other. However, hypernymy and subsumption are not at the same level of analysis since they belong to different fields of study and they describe and represent different kinds of realities: hypernymy relates lexical-semantic features between terms and may be captured by the linguist given their semantic field, or through the grammatical category they share in the syntagmatic axis (see Cruse, 2002). While subsumption, typically written as $C \sqsubseteq D$, has to do with formal languages for concept description, *"where the first concept always denotes a subset of the set denoted by the second one*" (Nardi & Brachmanv, 2003, p. 13). This particular relation plays a central role in Description Logics[121] since it "*is* [at]

[121] According to Baader and Nutt, it "is the most recent name for a family of knowledge representation (KR) formalisms that represent the knowledge of an application domain (the "world") by first defining the relevant concepts of the domain (its terminology), and then using these concepts to specify properties of objects and individuals occurring in the domain (the world description) […] Description Logics support inference patterns that occur in many applications of intelligent information processing systems, and which are also used by humans to structure and understand the world: classification of concepts and individuals. Classification of concepts determines subconcept/superconcept relationships (called subsumption relationships in DL) between the concepts of a given terminology, and thus allows one to structure the terminology in the form of a subsumption hierarchy. This hierarchy provides useful

the basic inference on concept expressions" (*ibid.*). Therefore, considering the disparate objects of study these two relations fall into, we shall make a terminological distinction between the different levels of analysis they belong to – where the former corresponds to the linguistic dimension and the latter to the conceptual dimension.

An identical isomorphic aspect is seen in the relations of meronymy vs. part-whole. The former belongs to the linguistic dimension and has several sub-types. The latter, in turn, is a *partitive*[122] *concept relation* (ISO/FDIS 1087, 2019 (E), p. 4), and has also its sub-types. We have noticed that the designation "part-whole" is widely used to paraphrase "meronymy" in the literature of different fields of study, namely in Lexicography, Cognitive Linguistics, and Terminology (see Neveu, 2015; Winston, Chaffin, & Hermann, 1987; Sager, 1990). Furthermore, we have observed that "meronymic" and "partonomic" relations are mentioned as synonymous in studies of computational knowledge representation (see Pribbenow, 2002). Nevertheless, and identically to the above-mentioned relations of hypernymy and subsumption, we make a terminological distinction between meronymy and partitive relations.

Therefore, when referring to the partitive relation, we will use the dichotomy [PART-WHOLE] in this study to describe the functionality of each compositional-element of a concept.

Finally, the associative[123] relation is uniquely conceptual and has a vast scope of sub-types given its pragmatic nature (see ISO/FDIS 1087, 2019 (E)). We have similarly marked the sub-types of the associative relation with a dichotomic label, depending on the pragmatic nature involved, e.g., [OBJECT-FUNCTION]. In the domain we are describing, where processes and activities are at the core of its technical nature, the

information on the connection between different concepts, and it can be used to speed-up other inference services." (2003, p. 47).

[122] The partitive relation is also named, in ontological works, as *partonomy*: "[…] the partitive hierarchy (partonomy) reflects the a priori part-whole relation between concepts […]". (Bernauer, 1994).

[123] designated as "complex relationship" by Sager (1990, p. 34).

associative relation is one of the most evidenced in our study, as shall be further demonstrated.

5.1. Conceptual analysis of Definition 1

The following conceptual analysis concerns Definition 1, a textual definition of <Stopper>, which we have first analysed linguistically (Section 4.2.1.). As mentioned before, we have first analysed the behaviour of lexical items in the syntagmatic axis to identify lexical-semantic relations established between terms, which, in turn, are related by the lexical marker that expresses the identified relation.

The conceptual analysis corresponds to the second stage of the analysis of Definition 1.

The observations of this analysis are systematised below in Table 23 and are based on the lexical markers found in the linguistic analysis of Definition 1. We must highlight that conceptual relations are not straightforwardly drawn from lexical-semantic relations. We have first developed conceptual relation identifiers based on

(i) the interpretation of the meaning (concept) pointed at by the terms in a set of two terms; and

(ii) the meaning pointed at by the linguistic expression underlying the lexical marker that relates those terms.

Table 23: Conceptual analysis of Definition 1: <Stopper>

CONCEPTUAL DIMENSION	**Aristotelian formula (X=Y+DC)** X [species] = Y [genus] + DC [differential characteristic]					
	Analysis	**Conceptual relation identifier**	**Conceptual relation**	**Interpretation**	**Transcription in X=Y+DC**	**Differential characteristics**
	stopper [is a] product	is_a [corresponds to LM 'is a']	**SUBSUMPTION**	stopper [SPECIES] product [GENUS]	stopper [SPECIES]= product [GENUS] + DC	
	stopper [is intended] to seal bottles or other containers	has_function [corresponds to LM 'intended to']	**ASSOCIATIVE**	stopper [OBJECT] to seal bottles [FUNCTION]	stopper [SPECIES] = product [GENUS] + to seal bottles [DC]	/to seal bottles/
	stopper [consists of] one piece	has_part [corresponds to 'consisting of']	**PARTITIVE**	stopper [WHOLE] one piece [PART]	stopper [SPECIES] = product [GENUS] + one piece [DC]	/one piece/
	stopper [consists of] more (=several) pieces	has_part [corresponds to 'consisting of']	**PARTITIVE**	stopper [WHOLE] several pieces [PART]	stopper [SPECIES] = product [GENUS] + several pieces [DC]	/several pieces/
	stopper [is obtained from] natural cork	has_raw_material [corresponds to 'obtained from']	**ASSOCIATIVE**	stopper [PRODUCT] natural cork [RAW MATERIAL]	stopper [SPECIES] = product [GENUS] + natural cork [DC]	/natural cork/

	stopper [is obtained from] natural cork and agglomerated cork	has_raw_material [corresponds to 'obtained from']	**ASSOCIATIVE**	stopper [PRODUCT] natural cork and agglomerated cork [RAW MATERIAL]	stopper [SPECIES] = product [GENUS] + natural cork and agglomerated cork [DC]	/natural cork and agglomerated cork/
	stopper [is obtained from] agglomerated cork	has_raw_material [corresponds to 'obtained from']	**ASSOCIATIVE**	stopper [PRODUCT] agglomerated cork [RAW MATERIAL]	stopper [SPECIES] = product [GENUS] + agglomerated cork [DC]	/agglomerated cork/
	stopper [is obtained from] natural cork	has_substance [corresponds to 'obtained from']	**ASSOCIATIVE**	cork [MATTER/SUBSTANCE] natural [PROPERTY]	natural cork [SPECIES] = cork [GENUS] + natural [DC]	/natural/
	stopper [is obtained from] agglomerated cork	has_substance [corresponds to 'obtained from']	**ASSOCIATIVE**	cork [MATTER/SUBSTANCE] agglomerated [PROPERTY]	natural cork [SPECIES] = cork [GENUS] + agglomerated [DC]	/agglomerated/

As we can observe above in Table 23, we have identified 9 conceptual relation identifiers based on the LM and the characteristics identified during the linguistic analysis. We inferred them through the meaning of the terms linked by the LM in addition to the information that is being pointed at. The rationale behind the inference is demonstrated in the following lines.

Sample 1: The conceptual relation identifier is_a and has_function, and the characteristic /to seal bottles/

Analysis	**Conceptual relation identifier**	**Conceptual relation**	**Interpretation**	**Transcription in X=Y+DC**	**Differential characteristics**
stopper [is a] product	is_a [corresponds to LM 'is a']	**SUBSUMPTION**	stopper [SPECIES] product [GENUS]	stopper [SPECIES]= product [GENUS] + DC	
stopper [is intended] to seal bottles or other containers	has_function [corresponds to LM 'intended to']	**ASSOCIATIVE**	stopper [OBJECT] to seal bottles [FUNCTION]	stopper [SPECIES] = product [GENUS] + to seal bottles [DC]	/to seal bottles/

Sample 1 above, represents the two first lines of Table 23.

As noted in the second example-line of Sample 1, we propose the conceptual relation identifier *has_function,* which corresponds to the lexical marker *"intended to"* referring to the function of the object. From the interpretation of this information, we assume that an associative conceptual relation is in place, a sub-type OBJECT-FUNCTION, in which *stopper* points to the meaning of OBJECT, and *to seal bottles* points to the meaning of FUNCTION. This interpretation can be represented as follows:

[stopper] OBJECT *has_function* [to seal bottles] FUNCTION

The dichotomy OBJECT-FUNCTION has a twofold importance at this point of the conceptual analysis: it underpins the sub-type of the associative relation, on the one hand, and enters, on the other, in the Aristotelian formula[124] known as X = Y + DC, where X=specific concept; Y=genus; and DC=differential characteristics. The purpose of using

[124] See Pearson (1998); Meyer (2001).

such formula aims at identifying the descriptive characteristics stated in the definition under analysis for the task of concept modelling.

To use such formula, one needs first to identify two concepts: the specific concept and its genus. For that, we must look at the first example-line under the head *Interpretation* in Sample 1. In the linguistic analysis, we have seen that the term "stopper" is the specific term, and "product" is the generic one. The meaning of the concept <Stopper> is more specific than the meaning of the concept <Product>, thus, the conceptual relation established between these two concepts is subsumption[125] – a hierarchical relation in which a given generic concept (genus) subsumes specific concepts (species). Hence, <Stopper> is the subordinate concept, which we labelled [SPIECES], and <Product> is the superordinate concept, in turn labelled [GENUS]. This assumption can be represented as:

[stopper] $_{\text{SPECIES}}$ *is_a* [product] $_{\text{GENUS}}$

Once identified the genus and the species, we can then insert these two elements in the formula X $_{\text{SPECIES}}$ = Y $_{\text{GENUS}}$ + DC, where:

X = stopper; Y = product

At this point, there are no differential characteristics, since we are describing the most generic concept. This means that the conceptual relation of subsumption is clearly set between <Stopper> and <Product>, but no other information is captured. Differential characteristics are found at a later stage.

[125] According to Johansson, "In complete Aristotelian definitions, one starts from the highest genus and presents, stepwise, the definitions of the lower classes until the lowest classes (species) have been defined. In each such step the subsuming class is divided into two or more subsumed classes by means of some quality or property requirements. The classic Aristotelian example is "man =def rational animal"; meaning that the subsumed class "man" is defined by means of a more general subsuming class ("animal") plus a quality requirement, namely that the class "man" should have the quality "rationality" as its specific difference in relation to the other classes on the same level." (2008, p. 243).

We will focus again on the second example-line of Sample 1, replicated below, with the conceptual relation identifier *has_function* to demonstrate how we have identified differential characteristics:

Analysis	**Conceptual relation identifier**	**Conceptual relation**	**Interpretation**	**Transcription in X=Y+DC**	**Differential characteristics**
stopper [is intended] to seal bottles or other containers	has_function [corresponds to LM 'intended to']	**ASSOCIATIVE**	stopper [OBJECT] to seal bottles [FUNCTION]	stopper [SPECIES] = product [GENUS] + to seal bottles [DC]	/to seal bottles/

The two pieces of information represented above can be summed up as:

(1) a *stopper* = OBJECT + *to seal bottles* = FUNCTION

(2) a *stopper* is a species (a kind) of *product.*

These two types of information allow the following interpretation: what differentiates this *product* from other kinds of products is the feature of having a function, which in turn is *to seal bottles.* The [FUNCTION] is, therefore, the differential characteristic we need to complete the formula $X_{SPECIES} = Y_{GENUS} + DC$.

Thus, knowing that

$_{SPECIES}$ = stopper, $_{GENUS}$ = product, and DC = to seal bottles,

the transcription into the formula is:

stopper [SPECIES] = product [GENUS] + to seal bottles [FUNCTION=DC]

which leads us to assert that one of the characteristics found in the definition under analysis is: **/to seal bottles/**, as recorded in the second example-line of Sample 1, under the heading "differential characteristics".

Sample 2 below represents the third and fourth lines of Table 23 (p. 170).

Sample 2: The differential characteristics DC = /one piece/ and DC = /several pieces/

Analysis	Conceptual relation identifier	Conceptual relation	Interpretation	Transcription in X=Y+DC	Differential characteristics
stopper [consists of] one piece	has_part [corresponds to 'consisting of']	**PARTITIVE**	stopper [WHOLE] one piece [PART]	stopper [SPECIES] = product [GENUS] + one piece [DC]	/one piece/
stopper [consists of] more (=several) pieces	has_part [corresponds to 'consisting of']	**PARTITIVE**	stopper [WHOLE] several pieces [PART]	stopper [SPECIES] = product [GENUS] + several pieces [DC]	/several pieces/

As we can see in Sample 2, the conceptual relation identifier *has_part* is under focus.

The conceptual relation identifier *has_part* corresponds to the meaning conveyed by the lexical marker *"consisting of",* which relates the two terms "one piece" and "several pieces". Since the information conveyed by these two terms carries the notion of *part*, we have proposed a conceptual relation identifier that coherently assists this notion, namely *has_part.* From the interpretation of this information, we conclude that the conceptual relation established between the concepts pointed by "stopper" and "pieces" is partitive:

[stopper] $_{\text{WHOLE}}$ *has_part* [several pieces] $_{\text{PART}}$

This means that we have finally identified the DC = several pieces. From here, we can step directly to the formula X $_{\text{SPECIES}}$ = Y $_{\text{GENUS}}$ + DC because we already know that stopper = SPECIES and product = GENUS.

Thus, knowing that

$_{\text{SPECIES}}$ = stopper, $_{\text{GENUS}}$ = product, and DC = several pieces,

the transcription into the formula is:

stopper [SPECIES] = product [GENUS] + several pieces [PART=DC]

The same applies to the information conveyed by the term “one piece”. Hence, two more characteristics were identified in Definition 1, namely /one piece/ and /several pieces/.

Sample 3 below, represents further lines extracted from Table 23 (p. 170).

Sample 3: The differential characteristics /natural cork/, /natural and agglomerated cork/ and /agglomerated cork/

Analysis	Conceptual relation identifier	Conceptual relation	Interpretation	Transcription in X=Y+DC	Differential characteristics
stopper [is obtained from] natural cork	has_raw_material [corresponds to ‘obtained from’]	**ASSOCIATIVE**	stopper [PRODUCT] natural cork [RAW MATERIAL]	stopper [SPECIES] = product [GENUS] + natural cork [DC]	/natural cork/
stopper [is obtained from] natural cork and agglomerated cork	has_raw_material [corresponds to ‘obtained from’]	**ASSOCIATIVE**	stopper [PRODUCT] natural cork and agglomerated cork [RAW MATERIAL]	stopper [SPECIES] = product [GENUS] + natural cork and agglomerated cork [DC]	/natural cork and agglomerated cork/
stopper [is obtained from] agglomerated cork	has_raw_material [corresponds to ‘obtained from’]	**ASSOCIATIVE**	stopper [PRODUCT] agglomerated cork [RAW MATERIAL]	stopper [SPECIES] = product [GENUS] + agglomerated cork [DC]	/agglomerated cork/

The conceptual relation identifier we mentioned in the three lines of Sample 3 is *has_raw_material.* The motivation for its elaboration is based on the meaning pointed at by the terms related by the lexical marker “obtained from”. As we know from the linguistic analysis, the term “natural cork” points at the notion of substance, a material that a given object can be made of. The same applies to “agglomerated cork”. Since <Stopper> is an object made of a substance – which, in turn, has three types of possible combinations – we propose the conceptual relation identifier *has_raw_material* to represent such semantic dependency. This semantic dependency mirrors a pragmatic association, in which a <Stopper> is a [PRODUCT] obtained from a substance, more specifically a [RAW MATERIAL]. Considering the three types of substance combination that we had previously identified in the linguistic analysis, we will represent this interpretation in 3 lines:

1. [stopper] PRODUCT *has_raw_material* [natural cork] RAW MATERIAL

2. [stopper] PRODUCT *has_raw_material* [agglomerated cork] RAW MATERIAL

3. [stopper] PRODUCT *has_raw_material* [natural cork and agglomerated cork] RAW MATERIAL

Once again, we can step directly to the formula X + Y = DC, for we already know the genus and the species concepts: stopper = SPECIES and product = GENUS. In this case, the DC = [RAW MATERIAL] and the transcription into the formula is:

1.1 stopper [SPECIES] = product [GENUS] + natural cork [RAW MATERIAL=DC]

1.2 stopper [SPECIES] = product [GENUS] + agglomerated cork [RAW MATERIAL=DC]

1.3 stopper [SPECIES] = product [GENUS] + natural cork and agglomerated cork [RAW MATERIAL=DC]

As demonstrated, three more characteristics were identified, namely /natural cork/, /agglomerated cork/ and /natural cork and agglomerated cork/.

Sample 4: Conceptual relation identifier has_substance and the characteristics /natural/ and /agglomerated/

Analysis	Conceptual relation identifier	Conceptual relation	Interpretation	Transcription in X=Y+DC	Differential characteristics
stopper [is obtained from] natural cork	has_substance [corresponds to 'obtained from']	**ASSOCIATIVE**	cork [MATTER/SUBSTANCE] natural [PROPERTY]	natural cork [SPECIES] = cork [GENUS] + natural [DC]	/natural/
stopper [is obtained from] agglomerated cork	has_substance [corresponds to 'obtained from']	**ASSOCIATIVE**	cork [MATTER/SUBSTANCE] agglomerated [PROPERTY]	natural cork [SPECIES] = cork [GENUS] + agglomerated [DC]	/agglomerated/

The last observations written in Table 23 (p. 170) are shown above in Sample 4. The concepts analysed are the same as shown in sample 3, as well as the lexical markers. The difference in this analysis is that the conceptual relation identifier is not based on the LM, but the meaning of the relations and the concepts.

Our goal here was to make a deeper analysis of the concepts pointed at by the terms "natural cork" and "agglomerated cork". Recalling the *determined* and the *determining* of Felber (1984), we believe that it is possible to apply here the author's

perspective: “natural cork” and “agglomerated cork” are a sort of *short definition.* From Felber’s perspective, we can analyse the two terms as follows:

cork + natural = natural cork
cork = *determined member (constituent)* – is the term pointing to the genus concept
and
natural = *determining member* – is the term pointing to the differential characteristic

To demonstrate the above interpretation, we proposed the conceptual relation identifier: *has_substance,* based on the meaning of substance pointed at by the concept designated by “natural cork”. Once again, we assume that an associative relation is established, given the meaning of the concepts pointed at by “natural” and “cork”. That is, while the latter points at the notion of [MATTER/SUBSTANCE], the former points at its [PROPERTY], therefore a pragmatic dependency is observed:

[cork] $_{\text{MATTER/SUBSTANCE}}$ *has_substance* [natural] $_{\text{PROPERTY}}$

Based on Felber’s model, and focusing on “cork” and “natural”, the transcription into the formula X $_{\text{SPECIES}}$ = Y $_{\text{GENUS}}$ + DC would be:

If cork = genus, and natural = DC, we can assume that

natural cork [SPECIES] = cork [GENUS] + natural [DC]

The same methodology applies to “agglomerated cork”. With this last analysis, we have demonstrated that /natural/ and /agglomerated/ are differential characteristics, although not focusing on lexical markers but following Felber’s model. As mentioned before, Felber’s model is an additional mechanism to infer descriptive characteristics. In our view, it is an interesting model for the analysis of polylexical terms, but not sufficient to analyse large textual definitions.

The methodology described above is the foundation of our analysis of definitions written in natural language. It is an iterative work involving several tasks and several steps, where each of these depends (or not) from the previous one, both in the linguistic and the conceptual dimensions. To achieve this analysis, we first interpret the meaning

each term is pointing at, along with the meaning of the lexical marker that intermediates the establishment of a lexical-semantic relation between those terms, so that we can infer special knowledge information. This information is conveyed by the types of lexical relations expressed by those lexical markers and the terms they relate in the syntagmatic axis. Secondly, it is the interpretation of these lexical relations that allows us to reach conceptual information, not in a straightforward manner but through several mechanisms, like the ones shown above for the conceptual relation identifiers *is_a* and *has_function.*

5.1.1. Function, parts and substance

In the following lines, we shall demonstrate the role of the conceptual relation identifiers we have described in the previous section for the representation of concepts.

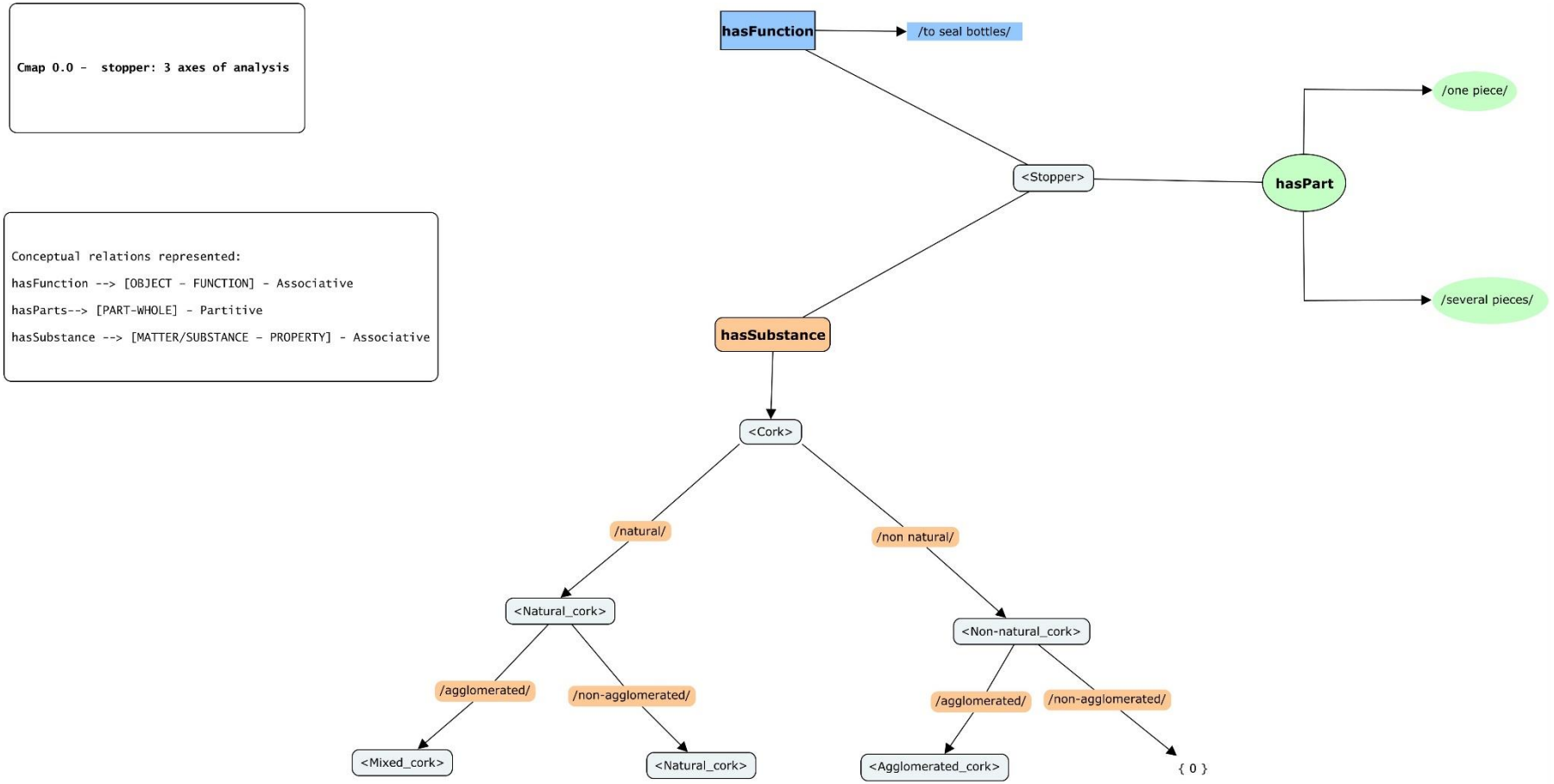


Figure 35: Cmap 0.0 – Representation of 3 axis of analysis: <Function>; <Parts> and <Substance> based on Definition 1

Cmap 0.0 is the representation in CmapTools[126] of 3 axes of analysis: function, parts and substance, based on the characteristics found in the linguistic analysis of Definition 1, in combination with the conceptual relation identifiers we have inferred during the conceptual analysis.

On this first map (Figure 35), we can observe the concept <Stopper> defined according to (1) Function, (2) Substance, and (3) Parts. These axes of analysis are at the basis of the conceptual relation identifiers we proposed, as a result of the analysis of Definition 1. We came up namely with has_function, has_substance and has_parts, which correspond respectively to hasFunction, hasSubstance and hasParts in our concept model. The role of these relationships is essential for the description of concepts, as we will demonstrate further on.

The relation henceforth written as hasFunction plays the role of a conceptual associative relation, subtype [OBJECT-FUNCTION]. Its graphical representation in Cmap 0.0 is modelled with horizontal straight arcs to avoid a representation in a hierarchical tree[127], a decision made according to (ISO 1087-1, 2000) recommendations. Therefore, to make a clear distinction between concepts that are modelled by associative relations from the ones systematised by subsumption (*genus-differentia)*, we decided for colour- and shape-node conventions: the associative relations are surrounded by a square-edge rectangular node, coloured in dark blue, while subsumed concepts are vertically represented in a round-edge rectangular node, coloured in grey.

[126] Software available online: https://cmap.ihmc.us/cmaptools/. CmapTools does not require graphical conventions; however, we decided, as a good practice, to label concepts with CamelBack notation within the different tools used throughout the study. In this environment, concepts are represented in nodes and underscored when several linguistic forms are involved, e.g., <Concept_1>; while characteristics are represented as arcs and written between forward slashes, e.g., /characteristic/. Furthermore, "Although there are no mandatory naming conventions for OWL classes, [it is recommended] that all class names should start with a capital letter and should not contain spaces. (This kind of notation is known as CamelBack notation)" (Horridge, 2011, p. 17).

[127] According to standards ISO 1087-1 (2000) and ISO 704 (2009), there is not a hierarchical dependency between concepts entertaining associative relations.

hasStructure is henceforth the designation we use for the conceptual partitive relation instead of "hasPart", as depicted in Conceptual Map 0.0. Here, concepts holding the relation "hasPart" are written in a green-coloured-oval node, and the partitive relation is modelled with square arcs. A convention in line with ISO 1087-1 (2000).

The relation hasStructure plays a key role in our study: it relates concepts carrying the meaning of "part" with concepts falling in the category of a "*composite concept*" (Bernauer, 1994, p. 2), for instance, <Disc> and <TechnicalStopper>. The classification of the former corresponds to the first element of the conceptual relation identifier [PART-WHOLE], whereas <TechnicalStopper> corresponds to the second element, i.e., <Disc>=is a part of <TechnicalStopper> and <TechnicalStopper>=is the whole.

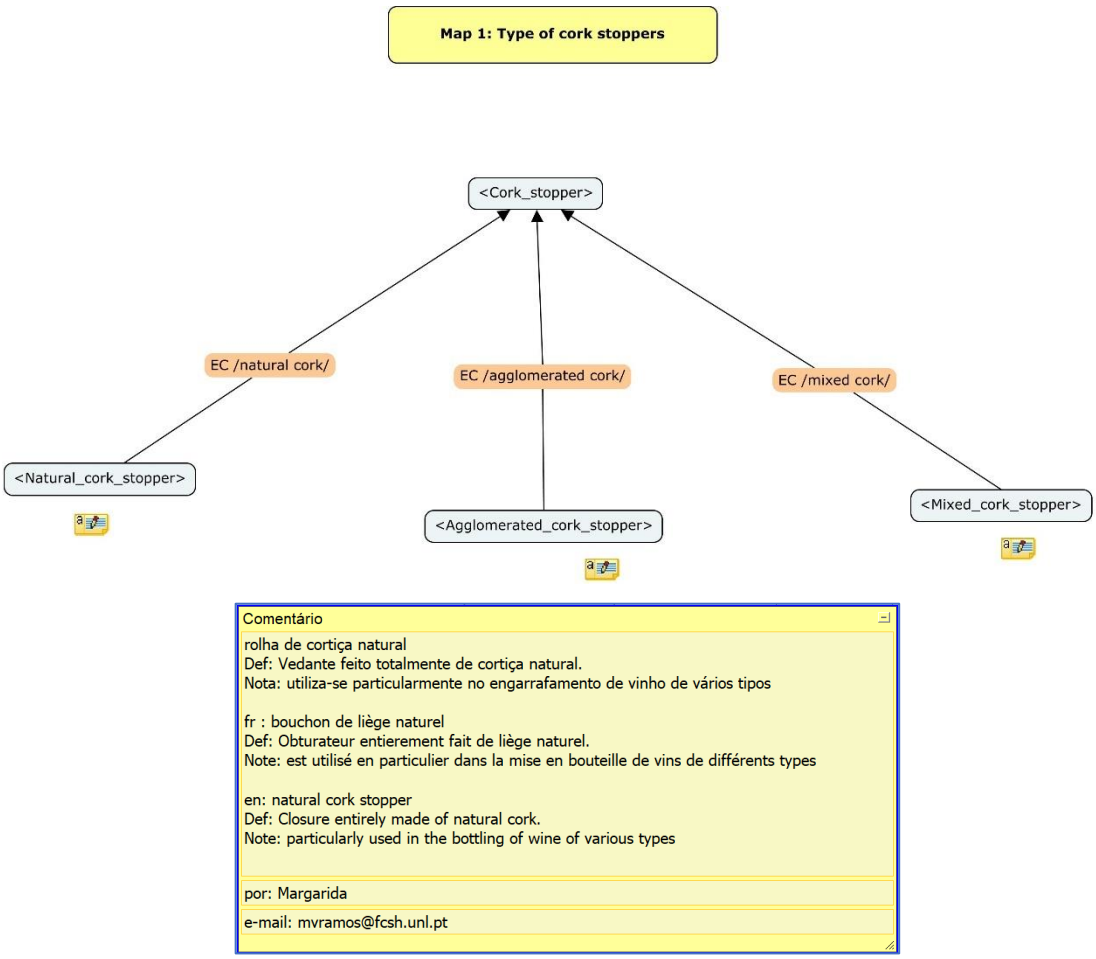
Finally, the relation designated hasSubstance plays the role of a conceptual associative relation, sub-type [MATTER/SUBSTANCE – PROPERTY]. Although not meant to represent a hierarchical relation, it is the root for the systematisation of the genus concept <Cork> and its *species* <Natural_cork>, <Non-natural_cork>, and so forth. The concept <Cork> is modelled through binary relations of *differentia*[128] with the guidance of three essential characteristics, namely /natural cork/, /non-natural cork/ and /agglomerated cork/, resulting then in a hierarchical-tree representation of a typology of cork. Here, characteristics are written in orange-coloured arcs. The purpose of this hierarchical-tree representation aims at demonstrating the role of essential characteristics[129] regarding the essence of cork, for it is *"indispensable to understanding a concept"* (ISO 1087-1, 2000, p. 3). In sum, cork is a raw material with three

[128] Base on the Isagoge epistemology, which relies on the principle of difference: "every difference added to something modifies it" [Isagoge 8.15-20] cit. by Roche (2012). The "principle of difference" was developed by Plato. Although a method initially criticised by Aristotle, the latter "admits, indeed advocates, division as a successful method of achieving (non-deductively), a definition. "Division is the only possible method," he says, "of avoiding the omission of any element of the essential nature." (Cassidy, 1967, p. 116).

[129] "is one of a set of characteristics that is both necessary and sufficient to determine the extension of a concept." (ISO 704, 2009, p. 2).

specifications, namely (1) natural (2) non-natural = agglomerated; and (3) mixed = natural and agglomerated.

This information can be represented in the form of a conceptual map:



Conceptual Map 1 – Representation of three types of cork stoppers based on the three types of raw material that a cork stopper can be made of represented in CmapTools.

Conceptual Map 1 is a map in which we represent three types of <Cork stoppers>, based on the three types of raw material that a cork stopper can be made of. Concepts are represented in nodes, and essential characteristics (EC) are represented as arcs.

As we can observe on Conceptual Map 1, we decided to label the genus concept as `<Cork_stopper>` instead of `<Stopper>`, for the sake of clarity. There is only one axis

of analysis, namely that of substance, for which the characteristics **/natural/**, **/agglomerated/** and **/mixed cork/** underlie the associative relationship IsMadeOf. This associative relation covers both sub-types [RAW MATERIAL-PRODUCT] and [MATTER/SUBSTANCE – PROPERTY]. In this extent, concepts are modelled according to the kind of substance they relate with, by virtue of the aforementioned essential characteristics, albeit with a differential role in the representation. Thus, the concept `<Cork_stopper>` subsumes 3 concepts (i.e., it has 3 species – also known as the extension of the concept): `<Natural_cork_stopper>`, `<Agglomerated_cork_stopper>` and `<Mixed_cork_stopper>`, in a co-relationship of siblings (also known as coordinated concepts).

Each of those sibling-concepts has an annotation with additional information (an icon, in yellow, placed near the node). When expanded, the user has access to its content, which is the definition of the concept in natural language; a feature that is applied by working in CmapTools. In this example, we can see the definition written in three languages: PT, FR, and EN.

One possible reading of Conceptual Map 1 is: an `<Agglomerated_cork_stopper>` isA `<Cork_Stopper>` and isMadeOf **/agglomerated cork/**.

5.2. Conceptual analysis of Definition 2

The following conceptual analysis will likewise focus on the definition of <Stopper>. As mentioned before, Definition 2 completes the information of Definition 1.

The conceptual analysis of Definition 2 was carried out following the same methodology stated for Definition 1: we have systematised our observations below in Table 24, in which the conceptual relations are based on conceptual relation identifiers, in turn, elaborated after the lexical markers previously identified in the linguistic analysis.

Table 24: Conceptual analysis of Definition 2

CONCEPTUAL DIMENSION	**Aristotelian formula (X=Y+DC)** X [species] = Y [genus] + DC [differential characteristic]					
	Analysis	**Conceptual relation identifiers**	**Conceptual relation**	**Interpretation**	**Transcription in X=Y+DC**	**Differential characteristics**
	stopper [is a] piece of cork	*is_a* [corresponds to LM 'is a']	**SUBSUMPTION**	stopper [SPECIES] piece of cork [GENUS]	stopper [SPECIES]= piece of cork [GENUS] + DC	
	stopper [is intended to] seal containers	*has_function* [corresponds to LM *'intended to'*]	**ASSOCIATIVE**	stopper [OBJECT] to seal containers [FUNCTION]	stopper [SPECIES] = piece of cork [GENUS] + to seal containers [FUNCTION=DC]	*/to seal containers/*
	stopper [is made of] cork	*has_substance* [correspondes to LM *'piece of'*]	**ASSOCIATIVE**	stopper [PRODUCT] cork [RAW MATERIAL]	stopper [SPECIES] = piece [GENUS] + cork [DC]	*/cork/*
	stopper [is] cylindrical	*has_shape* [corresponds to LM *'usually'*]	**ASSOCIATIVE**	stopper [OBJECT] cylindrical [SHAPE]	stopper [SPECIES] = piece of cork [GENUS] + cylindrical [DC]	*/cylindrical/*
	stopper [is] frustoconical	*has_shape* [corresponds to LM *'usually'*]	**ASSOCIATIVE**	stopper [OBJECT] conical [SHAPE]	stopper [SPECIES] = piece of cork [GENUS] + conical [DC]	*/ frustoconical /*
	stopper [is] prismatic quadrangular	*has_shape* [corresponds to LM *'usually'*]	**ASSOCIATIVE**	stopper [OBJECT] prismatic quadrangular [SHAPE]	stopper [SPECIES] = piece of cork [GENUS] + prismatic quadrangular [DC]	*/ prismatic quadrangular /*

	stopper [with] rounded lateral edges	*has_process* [corresponds to LM *'sometimes with'*]	**ASSOCIATIVE**	? = [PROCESS] rounded edges = [RESULT]	stopper [SPECIES] = piece of cork [GENUS] + rounded edges [DC]	*/rounded edges/*
	stopper [with] chamfered lateral edges	*has_process* [corresponds to LM *'sometimes with'*]	**ASSOCIATIVE**	? = [PROCESS] chamfered edges = [RESULT]	stopper [SPECIES] = piece of cork [GENUS] + chamfered edges [DC]	*/chamfered edges/*
	stopper [consists of] one element	*has_part* [corresponds to *'consisting of'*]	**PARTITIVE**	stopper [WHOLE] one element [PART]	stopper [SPECIES] = piece of cork [GENUS] + one element [DC]	*/one element/*
	stopper [consists of] several elements	*has_part* [corresponds to *'consisting of'*]	**PARTITIVE**	stopper [WHOLE] several elements [PART]	stopper [SPECIES] = piece of cork [GENUS] + several elements [DC]	*/several elements/*

As shown in Table 24, we have proposed 9 conceptual relation identifiers.

We will not address here the conceptual relation identifiers that were already discussed in the conceptual analysis of Definition 1, namely *is_a, has_function,* and *has_part* since the identified conceptual relations are identical, as well as the characteristics.

In the linguistic analysis (Section 4.2.2, p. 142), we have observed that the lexical-semantic relation of hyponymy does not always correspond to the conceptual relation of subsumption. The information captured through the lexical marker “usually” in fact relates to a specification; however, once in the conceptual dimension, the process of capturing differential characteristics required a finer granularity: we had to propose labels, each of these pointing at a conceptual field, under which the concepts designated by the terms are assigned, e.g., stopper = [OBJECT] and cylindrical = [SHAPE].

Finally, based on these labels, we were able to propose conceptual relation identifiers, e.g., *has_shape.* This method led us to observe that instead of a simple hierarchical relation of inclusion – subsumption – associative relations are more accurate to mirror the specificity of each concept given the more complex relationships involved. Thus, the dichotomy [OBJECT-SHAPE] is used to describe the associative dependency relation between concepts, where differential characteristics falling in the category of [SHAPE] are what determines the specificity of the concept falling in the category of [OBJECT].

The following lines will demonstrate the method we have just described.

Sample 5 below represents the two first lines of Table 24 – Conceptual analysis of Definition 2.

Sample 5: Sample of the Conceptual dimension analysis, retrieved from Table 24

Analysis	Conceptual relation identifier	Conceptual relation	Interpretation	Transcription in X=Y+DC	Differential characteristics
stopper [is] cylindrical	*has_shape* [corresponds to LM *'usually'*]	**ASSOCIATIVE**	stopper [OBJECT] cylindrical [SHAPE]	stopper [SPECIES] = piece of cork [GENUS] + cylindrical [DC]	*/cylindrical/*
stopper [with] rounded lateral edges	*has_process* [corresponds to LM *'sometimes with'*]	**ASSOCIATIVE**	? = [PROCESS] rounded edges = [RESULT]	stopper [SPECIES] = piece of cork [GENUS] + rounded edges [DC]	*/rounded edges/*

Looking at the first line-example of Sample 5, we have elaborated the conceptual relation identifier *has_shape* to mirror the meaning of stopper=[OBJECT] and cylindrical=[SHAPE]. This semantical dependency corresponds to the existence of an associative conceptual relation if we take into consideration that "[s]ome associative relations exist when dependence is established between concepts concerning their proximity in space or time. These relations may involve:

concrete item – shape" (ISO 704, 2009, p. 17).

Since the concept designated by the term "stopper" falls under the generic meaning of an object and the concept designated by "cylindrical", under shape, we can deduct that these two concepts are dependent in semantical proximity to which we assign the dichotomy [OBJECT-SHAPE]. This associative conceptual relation can be represented as:

[stopper] $_{\text{OBJECT}}$ *has_shape* [cylindrical] $_{\text{SHAPE}}$

From this interpretation, we can move to the Aristotelian formula X $_{\text{SPECIES}}$ = Y $_{\text{GENUS}}$ + DC.

Knowing that

$_{\text{SPECIES}}$ = stopper and $_{\text{GENUS}}$ = piece of cork,

cylindrical is the descriptive characteristic.

The transcription of this interpretation onto the formula is:

stopper [SPECIES] = piece of cork [GENUS] + cylindrical [DC]

For the second line-example in Sample 5, a similar decision favoured the associative relation instead of subsumption. In this case, the conceptual relation identifier *has_process* intends to mirror the associative relation underlying the dichotomy [PROCESS-RESULT], where differential characteristics falling in the category of [RESULT] are what determines the specificity of the genus concept, e.g., rounded edges = [RESULT].

The motivation behind the identifier *has_process* is grounded on a more in-depth morphosyntactic analysis of "rounded edges" in the linguistic dimension. The analysis of this adjectival structure, namely [Adj + N], in which the adjective morphologically derives from the Past Participle of the verb "to round", led us to infer that "rounded edges" are edges that were rounded off, pointing thus to an action that occurred in the past, i.e., the round-shape of "edges" is the result of a PROCESS, which we represent as follows:

[rounded edges] RESULT *has_process* [?] PROCESS

The question mark is due to the inexistence of a name to designate the involved PROCESS. The expert did not include this information in the textual Definition 3.

As demonstrated on the previous paragraph, despite the former analysis already undergone in the linguistic dimension, we had to revisit it to obtain information that had not been previously captured by any lexical marker: it was necessary to obtain information on the third stage by analysing the morphosyntactic structures of the terms – similarly to what we demonstrated with Felber's model (Section 4.1.2., p. 127). This back and forward is a common feature in terminological work as pointed out by Costa and Silva (2008): in the terminological workflow, the terminologist starts with texts whereby knowledge is reached, but at a given moment he/she returns to the text to stabilise knowledge. In our case, going back to the text had the primary purpose of

getting a more in-depth linguist analysis, since specific knowledge information is more likely to be obtained.

In sum, most of the novel differential characteristics found in Definition 2 fall under the conceptual markers [SHAPE] or [RESULT], e.g.:

stopper [SPECIES]= piece of cork [GENUS] + *frustoconical* = [SHAPE=DC]

stopper [SPECIES]= piece of cork [GENUS] + *chamfered edges* = [RESULT=DC]

The dichotomies [OBJECT-SHAPE] and [PROCESS-RESULT] have a double importance too: on the one hand, they represent differential characteristics to guide us through the systematisation of the concepts of the domain; and on the other, they correspond to two axes of analysis for the task of modelling concepts.

5.2.1. Complementary information found in Definition 2

Similarly to Definition 1, we shall demonstrate in the following lines the role played by the conceptual relation identifiers we have described in the previous section for the task of modelling concepts.

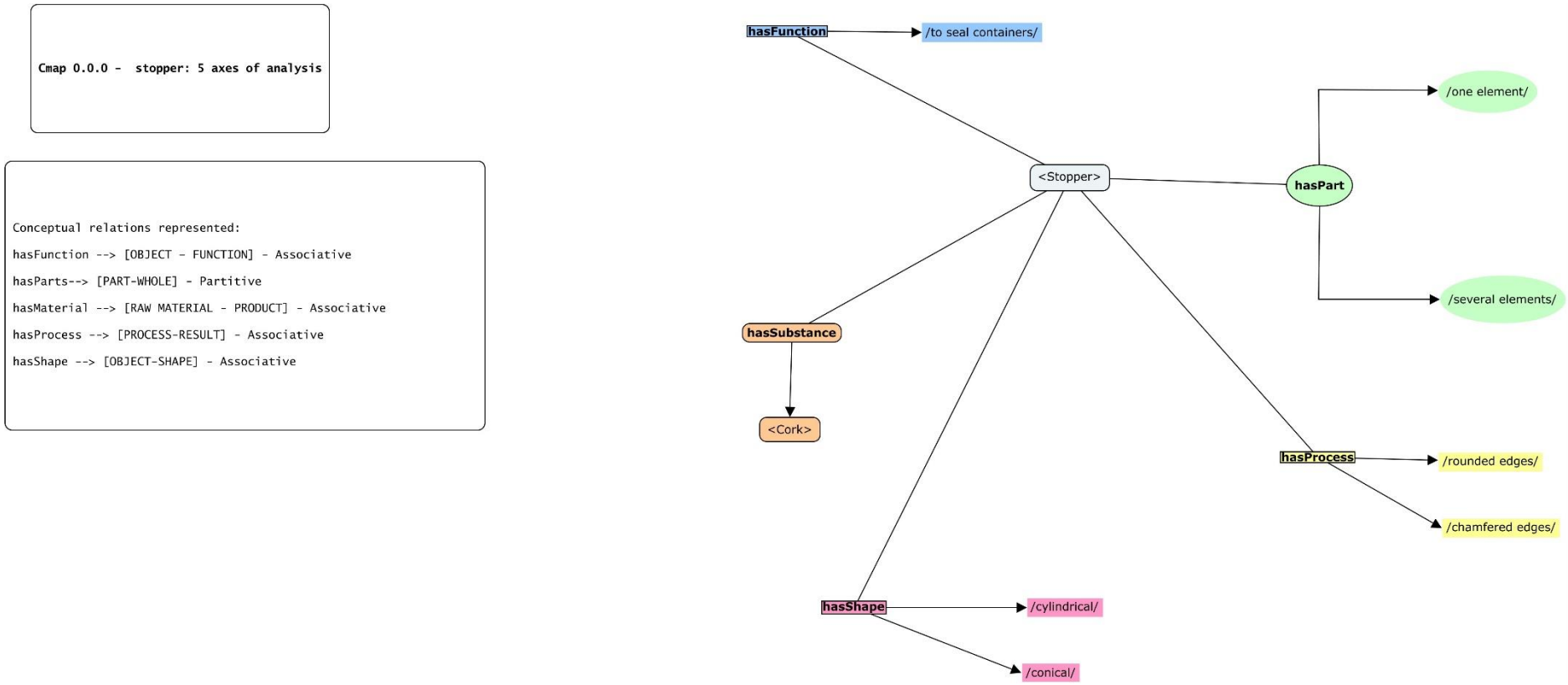


Figure 36: CMap 0.0.0 is an evolution of Cmap 0.0 with the addition of 2 axes of analysis: <Process> and <Shape>

Cmap 0.0.0 is an evolution of Cmap 0.0.

Based on Definition 2, two more axes of analysis were added, namely process and shape. These additional axes are based on the conceptual relation identifiers *has_process* and *has_shape*. Hence, the concept <Stopper> can be defined taking into consideration 5 axes of analysis: (1) Function, (2) Substance, (3) Parts, (4) FinishingProcess[130] and (5) Shape.

The relation henceforth named hasShape plays the role of a conceptual associative relation, sub-type [OBJECT-SHAPE]. Thus, concepts linked through the associative relation underlying this dichotomy are systematised by virtue of differential characteristics such as /conical/. According to ISO 1087 (2000) recommendations, the graphical representation of associative relations is modelled with horizontal straight arcs given their non-hierarchical feature.

Likewise modelled with horizontal straight arcs, we can see in Cmap 0.0.0 the relation henceforth named hasProcess; another associative relation, but within the sub-type [PROCESS-RESULT]. Here, concepts are systematised by resorting to certain characteristics, such as /with process X/, in which "X" is a kind of <FinishingProcess>, e.g., the process of <EdgeChamferingOperation>.

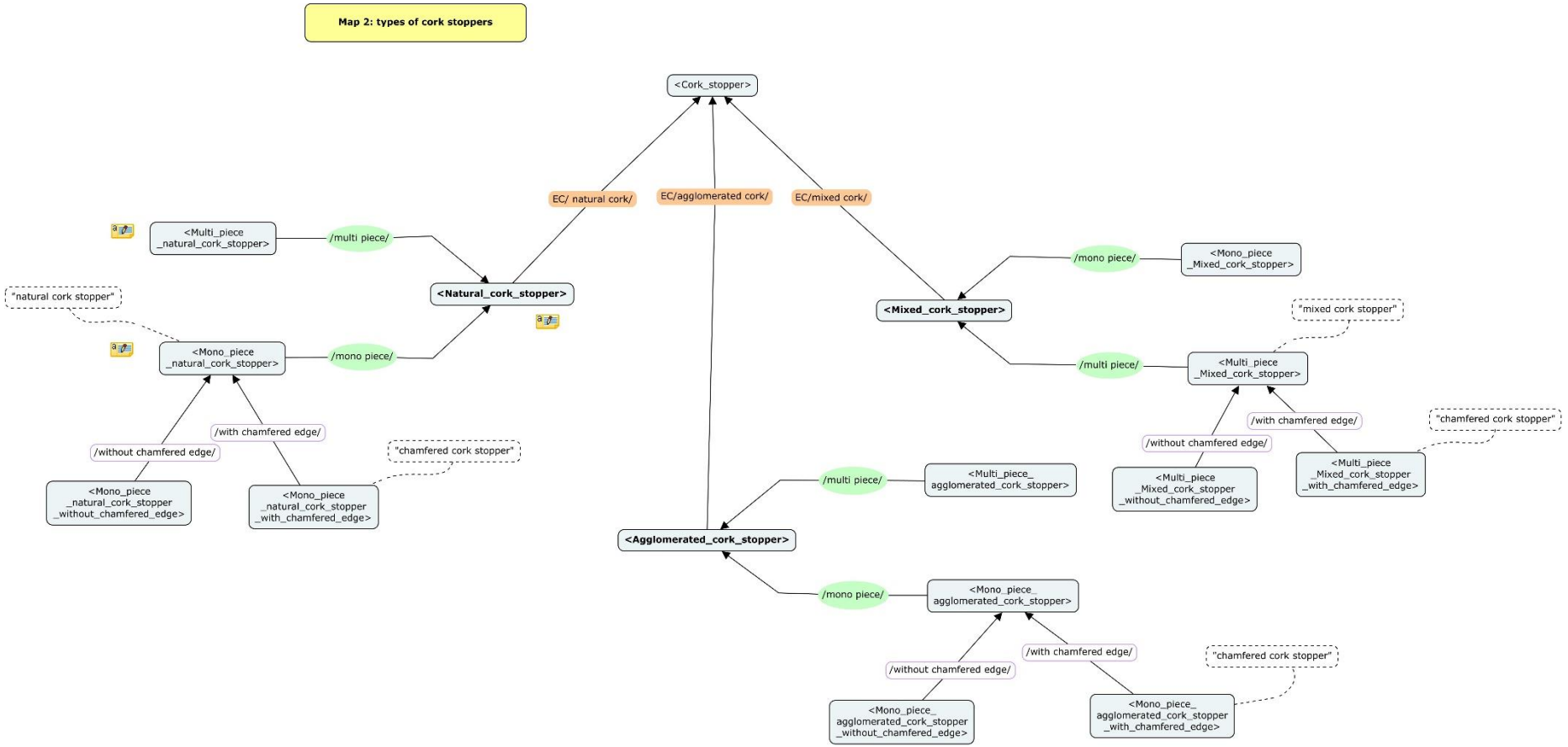
At this point, we need all the underpinning axes of analysis to model most of the extension[131] of <CorkStopper>. When we refer to "most of", we mean that the typology

[130] The concept of <FinishingProcesses> is a broad concept named after the definitions for the operations <EdgeChamferingOperation> and <EdgeRoundingOperation>. We must note that most of the "operations" involved are not referred by the expert in the definitions we have drawn from the corpus for cork stopper typology. Instead, definitions describe the substance or the structural composition of the stopper. Therefore, it was necessary to compile a different collection of definitions to systematise a typology of cork stoppers that are submitted to "finishing processes". Furthermore, a collection of definitions for a typology of operations had to be simultaneously compiled. Hence, in the manufacture of cork stoppers, depending on a given (or several) "finishing treatment" put in place, a specific cork stopper results, as further demonstrated in Map 2 and OntoGraf 3 (Section 6 – Building the ontology).

[131] in the sense that "the extensional definition of a generic or comprehensive concept consists in enumerating all its subordinate concepts (specific in the case of a generic concept, partitive for a comprehensive concept). This type of definition must not be confused with the "extensional definition" of a set (respectively to a concept) in mathematics which consists of enumerating the objects comprising that set (respectively belonging to the concept)." (Roche, 2012, p. 26).

of cork stoppers represented in this study is not exhaustive. Its completeness, though, is worthy of future work.

The following representation is a proposal of a conceptual representation in the form of a map in CmapTools, where we take into consideration differential characteristics drawn from the analysis of Definition 2, excluding /to seal containers/.



Conceptual Map 2 – Representation of Definition 1 and 2 in CmapTools, taking into consideration three axes of analysis: Substance, Parts and Finishing Process.

Map 2 is a conceptual representation of Definition 2 in CmapTools. Only three axes of analysis are considered: Substance, Parts and Finishing process.

Characteristics are highlighted in different colours to make a clear distinction between the sort of conceptual relationships they underlie. In this line, the characteristics /natural cork/, /agglomerated cork/ and /mixed cork/ are coloured in orange and underlie the systematisation of concepts holding the associative relationship IsMadeOf. The characteristics /mono piece/ and /multi-piece/ are coloured in green and underlie the division of concepts denoting constituent Parts. Finally, for the associative relation hasProcess, the underlying characteristics are surrounded by a purple line, e.g., /with chamfered edges/[132].

Within the axis of analysis of Finishing process, only one characteristic is considered, that of /with chamfered edges/ for economy of space. It would be necessary to create another map for the representation of these three types of cork stoppers through *differentia* dichotomy[133] using /with rounded edges/ or /without rounded edges/, given the three types of substance that a stopper might be made of.

Thus, one possible reading of Map 2 is: a `<Mono_piece_natural_cork_stopper_with_chamfered_edge>` is the specialisation of a `<Mono_piece_natural_cork_stopper>` that was submitted to `<EdgeChamferingOperation>`, which is a kind of `<FinishingProcess>`.

5.3. Conceptual analysis of Definition 3

The conceptual analysis of Definition 3 – a definition of <Natural cork stopper> – was carried out following the same methodology stated for Definitions 1 and 2.

---

[132] Despite its non-hierarchical nature, the associative relation hasProcess is modelled vertically to facilitate its systematisation gracefully.

[133] is the Aristotelian methodology described by its follower Porphyre. According to Spies and Roche, "the concept tree corresponding to an Aristotelian approach is usually binary since the very notion of difference in its traditional philosophical sense allows only dichotomous alternatives." (2006, p. 64).

As specified before, conceptual relations are not straightforwardly drawn from lexical-semantic relations. We first develop conceptual relation identifiers based on the interpretation of the meaning pointed by lexical markers.

Our observations are systematised below in Table 25.

Table 25: Conceptual analysis of Definition 3: <Natural cork stopper>

CONCEPTUAL DIMENSION	**Aristotelian formula (X=Y+DC)** X [species] = Y [genus] + DC [differential characteristic]					
	Analysis	**Conceptual relation identifier**	**Conceptual relation**	**Interpretation**	**Transcription in X=Y+DC**	**Differential characteristics**
	natural cork stopper [is a] stopper	*is_a* [corresponds to LM *'is a'*]	**SUBSUMPTION**	stopper [GENUS] natural cork stopper [SPECIES]	natural cork stopper [SPECIES] = stopper [GENUS] + DC ?	
	natural cork stopper [is made of] natural cork	*has_substance* [corresponds to LM *'consisting entirely of'*]	**ASSOCIATIVE**	natural cork stopper [PRODUCT] natural cork [RAW MATERIAL]	natural cork stopper [SPECIES] = stopper [GENUS] + natural cork [DC]	*/natural cork/*
	natural cork stopper [is made of] natural cork	*has_substance* [corresponds to LM *'consisting entirely of'*]	**ASSOCIATIVE**	cork [MATTER] natural [PROPERTY]	natural cork [GENUS] = cork [GENUS] + natural [DC]	*/natural/*
	natural cork stopper [is submitted to] sealing operation	*has_process* [corresponds to LM *'submitted to'*]	**ASSOCIATIVE**	sealing operation = [PROCESS] ? = [RESULT]	? [SPECIES] = natural cork stopper [GENUS] + sealing operation [DC]	*/sealing operation/*
	colmated natural stopper [is a] natural cork stopper	*is_a* [corresponds to the LM *'commonly referred as'*]	**SUBSUMPTION**	natural cork stopper [GENUS] colmated natural stopper [SPECIES]	colmated natural stopper [SPECIES] = natural cork stopper [GENUS] + colmated [DC]	*/colmated/*

The conceptual relations inscribed in Table 25 are based on the conceptual relation identifiers we have described after the lexical markers previously identified in the linguistic analysis of Definition 3. We will not address the conceptual relation identifier *is_a* (in the first line of Table 3) and *has_substance* for they are identical to the ones we have previously discussed in both Definition 1 and Definition 2.

The descriptive characteristics found in Definition 3 are also identical to what we have observed so far, namely /natural cork/, /natural/, except for two characteristics: /colmated/ and /sealing operation/.

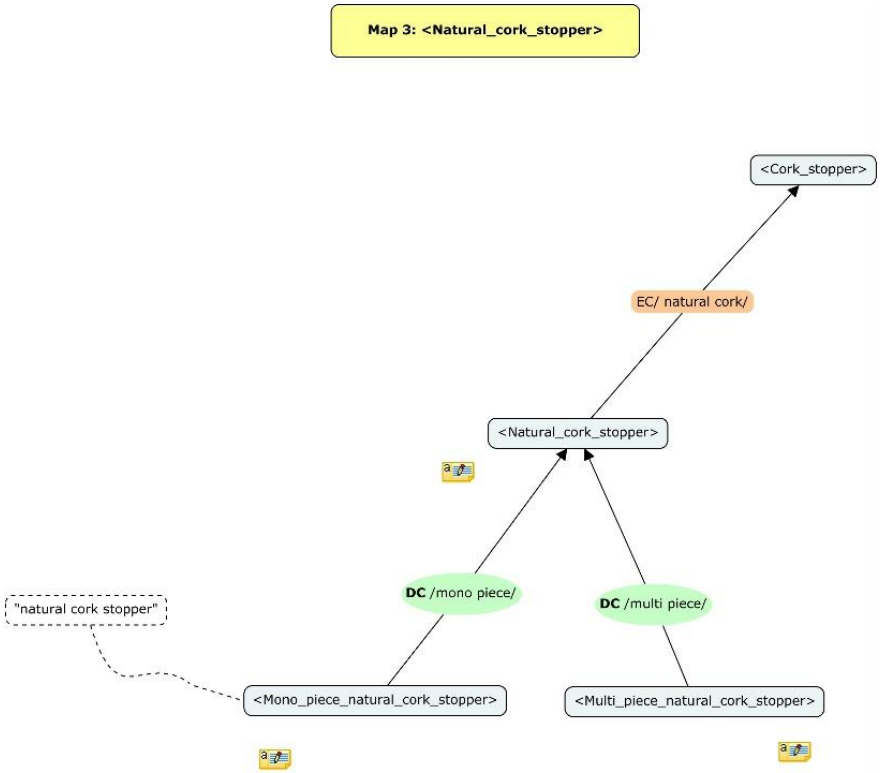
Although not discussed in this section, we must recall that a <Natural cork stopper> *is_a* <Stopper>, in which the hierarchical relation of subsumption is observed:

[natural cork stopper] SPECIES *is_a* [stopper] GENUS

and that a <Stopper> *has_part* either mono- or multipiece – two pieces of information that we obtained from the previous definitions / analyses:

[stopper] WHOLE *has_part* [several pieces] PARTS

The first statement of Definition 3 only conveys the information represented by the first interpretation above, where the dichotomy [SPECIES-GENUS] is identified. However, we will include the two pieces of information represented above in the next conceptual map.



Conceptual Map 3 – Two structures of <Natural_cork_stopper> in CmapTools

Conceptual Map 3 is the conceptual representation in CmapTools of the first statement of Definition 3, from which we have inferred that a <Natural cork stopper> *is_a* <Cork stopper>.

In this map, two axes of analysis are considered: Substance and Parts. The characteristic **/natural cork/** is the arc coloured in orange and underlies the associative relation isMadeOf; the characteristics **/mono piece/** and **/multi piece/** are the arcs coloured in green, and underlie the subdivision of concepts according to their number of Parts, i.e., one or several.

At the bottom of the map, there is a concept named <Mono_piece_natural_cork_stopper>. This concept name is the identification of one of the two specifications of <Natural_cork_stopper>, regarding its compositional parts. We would like to highlight that this form of concept identification intends to mirror the differential characteristics that promote the specialisation of the *genus proximum* concept. This reflexion is represented below:

(1) <Natural_cork_stopper> + /mono piece/ = <Mono_piece_natural_cork_stopper>

(2) <Natural_cork_stopper> + /multi piece/ = <Multi_piece_natural_cork_stopper>

As depicted in Conceptual Map 3, there is a dashed-line balloon attached to the node of the concept <Mono_piece_natural_cork_stopper>. This balloon represents the term – the verbal designation – of the concept to which it is attached. The option of inserting natural language items in an environment for conceptual representation – two different dimensions that should not be mixed up – intends to demonstrate how different the names of the concepts are when compared to the terms that designate them – in the sense of conceptualisation *versus* verbalisation. While the former is a sort of summary of characteristics, the latter is the linguistic form used in a specialised context of discourse by the expert of the domain. Additionally, the interpretation of the conceptual map is facilitated with verbal designations attached, which is (1) a helpful option for non-experts of the domain; and (2) at a given point, the identification of concepts tends to get very long, which culminates in an unfriendly reading.

The conceptual information represented in Conceptual Map 3, namely the axes of analysis Substance and Parts and underlying characteristics: /natural cork/ , /mono piece/ and /multi piece/ will be some of the coordinates for the elaboration of the formal description of the concept `<NaturalCorkStopper>` in Protégé (see Section 6.2.1, p. 227).

Finally, <Multi_piece_natural_cork_stopper> will help us to formally describe types of <Stoppers> composed of several Parts – not only made of <Natural_cork>, but also <Agglomerated_cork> and <Mixed_cork>. Here, the characteristics fall under the axis of analysis Parts and are the coordinates to model multi-part concepts. This subject will be addressed in more detail in Section 6.5, p. 275.

We can finally step into the second statement of Definition 3, namely the footnote from which we obtained the information: <Natural cork stopper> is submitted to /sealing operation/.

Sample 6: Conceptual relation identifier has_process and characteristic /sealing operation/

Analysis	**Conceptual relation identifier**	**Conceptual relation**	**Interpretation**	**Transcription in X=Y+DC**	**Differential characteristics**
natural cork stopper [is submitted to] sealing operation	*has_process* [corresponds to LM *'submitted to'*]	**ASSOCIATIVE**	sealing operation = [PROCESS] ? = [RESULT]	? [SPECIES] = natural cork stopper [GENUS] + sealing operation [DC]	*/sealing operation/*

Sample 6 represents the fourth line of Table 25 – Conceptual analysis of Definition 3.

As shown in Sample 6, the conceptual relation identifier *has_process* corresponds to the lexical marker "submitted to". Like the LM – that clearly expresses an action – the conceptual relation identifier intends to mirror the semantic dependency established between <Natural cork stopper> and the operation of sealing. The semantic dependency here observed falls under the associative relation of [PROCESS-RESULT] since "sealing operation" points to a process. However, <Natural cork stopper> does not point to a result, but a substance that undergoes a process:

[?] $_{RESULT}$ *has_process* [sealing operation] $_{PROCESS}$

This explains the incomplete information on the above representation. The same happens when transcribing this incomplete information into the Aristotelian formula X $_{SPECIES}$ = Y $_{GENUS}$ + DC. If <Natural cork stopper> is not the result, it means that it is not the most specific concept, therefore, it replaces the value of Y = genus in the Aristotelian formula:

$_{SPECIES}$ = ?, $_{GENUS}$ = natural cork stopper, and /sealing operation/ = DC

The meaning of the RESULT was inferred at a later moment, with the information conveyed by the characteristic /colmated/.

The novel differential characteristic /colmated/ was obtained after the interpretation of the analysis made on the [GENUS-SPECIES] relation held between the concepts designated by "colmated natural stopper" and "natural cork stopper". This conceptual relation was inferred from the lexical marker "commonly referred as", which points to a specification between the concepts pointed at by the above terms. Within this rationale, the first concept

designated by “colmated natural stopper” is more specific than the concept designated by “natural cork stopper”. This interpretation is represented as:

[colmated natural stopper] $_{SPECIES}$ *is_a* [natural cork stopper] $_{GENUS}$

and further transcribed into the Aristotelian formula X $_{SPECIES}$ = Y $_{GENUS}$ + DC:

knowing that

$_{SPECIES}$ = colmated natural cork stopper and $_{GENUS}$ = natural cork stopper,

we obtain the following reasoning:

colmated natural stopper [SPECIES] = natural cork stopper [GENUS] + colmated [DC]

The transcription of this conceptual relation into the Aristotelian formula clearly demonstrates a semantic dependency established though the generic-specific relation between these two concepts. The semantic dependency is assigned by the additional characteristic added to the more generic concept, namely the differential characteristic /colmated/ – a characteristic that determines the specification of one concept from another.

Finally, from the knowledge that a

[colmated natural stopper] $_{SPECIES}$ *is_a* [natural cork stopper] $_{GENUS}$

we can revisit and complete the previous analysis, where the concept pointing to a RESULT was missing:

knowing that

$_{SPECIES}$ = colmated natural stopper and $_{GENUS}$ = natural cork stopper,

we obtain:

colmated natural stopper [SPECIES] = natural cork stopper [GENUS] + sealing operation [DC]

which finally allows us to infer that:

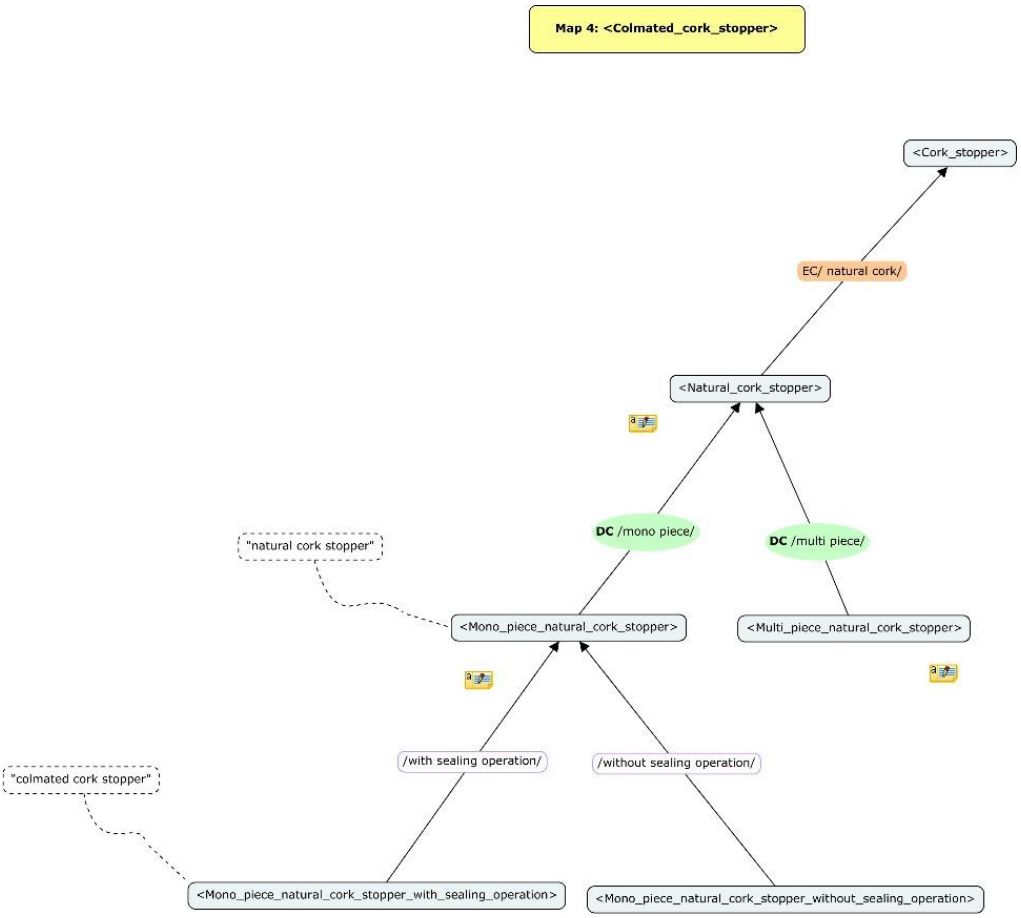
[colmated natural stopper] $_{RESULT}$ *has_process* [sealing operation] $_{PROCESS}$

As demonstrated above, the conceptual analysis does not follow a rigid model, in the sense of pursuing the same order of words used in discourse. The same was demonstrated during the linguistic analysis: some meanings are construed by the analysis of co-occurrences, where we acquire complementarity information.

The following conceptual map is the representation of the two sentences of Definition 3.

Recalling Conceptual Map 3 (p. 199), where the first statement of Definition 3 was represented by two axes of analysis, namely Substance and Parts, Conceptual Map 4 is an evolution of that preceding map.

Since Conceptual Map 4 is the conceptual representation of the second definitional statement of Definition 3, three axes of analysis are considered: Substance, Parts, and Finishing processes, to which the characteristics /with sealing operation/ and /without sealing operation/ surrounded in purple were added.



Conceptual Map 4 – Conceptual map of <Mono_piece_natural_cork_stopper_with_sealing_operation> in CmapTools

As we can see on Conceptual Map 4, the above characteristics led us to a different level of concept representation, i.e., the concept `<Mono_piece_natural_cork_stopper_with_sealing_operation>`, verbally designated by “colmated cork stopper”, is a specialisation of `<Mono_piece_natural_cork_stopper>`, in turn, verbally designated by “natural cork stopper”. Therefore, these two concepts should not be treated at the same level, nor defined within the same definitional context, either in natural language or (semi)formal languages.

5.4. Conceptual analysis of Definition 4

The methodology for the conceptual analysis of Definition 4 – a definition of the concept <Colmated natural cork stopper> – was followed in the same way we have demonstrated for Definitions 1, 2 and 3.

The observation of the conceptual analyses, which are based on the linguistic analyses that had been previously demonstrated, were recorded in Table 26.

Table 26: Conceptual analysis of Definition 4: <Colmated natural cork stopper>

	Aristotelian formula (X=Y+DC) X [species] = Y [genus] + DC [differential characteristic]					
	Analysis	**Conceptual relation identifier**	**Conceptual relation**	**Interpretation**	**Transcription in X=Y+DC**	**Differential characteristics**
CONCEPTUAL DIMENSION	colmated natural cork stopper [is a] stopper	*is_a* [corresponds to LM *'is a'*]	**SUBSUMPTION**	stopper [GENUS] colmated natural cork stopper [SPECIES]	colmated natural cork stopper [SPECIES] = stopper [GENUS] + [DC] ?	
	colmated natural cork stopper [is made of] natural cork	*has_raw_material* [corresponds to LM *'is made of'*]	**ASSOCIATIVE**	colmated natural cork stopper [PRODUCT] natural cork [RAW MATERIAL]	colmated natural cork stopper [SPECIES] = stopper [GENUS] + natural cork [DC]	*/natural cork/*
	colmated natural cork stopper [is made of] natural cork	*has_substance* [corresponds to LM *'is made of'*]	**ASSOCIATIVE**	natural cork [MATTER/SUBSTANCE] colmated [PROPERTY]	colmated natural cork stopper [SPECIES] = natural cork stopper [GENUS] + colmated [DC]	*/colmated/*
	colmated natural cork stopper whose [lenticels are filled] with cork powder	*has_process* [corresponds to LM *'lenticels are filled'*	**ASSOCIATIVE**	colmated natural cork stopper [RESULT] lenticels are filled [PROCESS]	colmated natural cork stopper [SPECIES] = natural cork stopper [GENUS] + filled lenticels [DC]	*/ filled lenticels /*

	cork powder [results from] the dimensional finishing processes of natural cork stoppers	*has_process* [corresponds to LM *'results from'*]	**ASSOCIATIVE**	cork powder [RESULT] dimensional finishing processes [PROCESS]	cork powder [SPECIES] = natural cork [GENUS] + dimensional finishing process [DC]	*/dimensional finishing process/*

The conceptual analysis of Definition 4 provided us with four characteristics, namely /natural cork/, /colmated/, / filled lenticels / and /dimensional finishing process/.

Similarly to the linguistic analysis of Definition 4, there is not much information to add to what had already been observed during the analysis of Definition 3, except for the two characteristics /filled lenticels/ and /dimensional finishing process/.

The first of the two characteristics above, namely /filled lenticels/, was inferred after the elaboration of the conceptual relation identifier *has_process,* which intends to mirror its corresponding lexical marker "lenticels are filled".

Sample 7: Conceptual relation identifier has_process and characteristic /filled lenticels/

Analysis	**Conceptual relation identifier**	**Conceptual relation**	**Interpretation**	**Transcription in X=Y+DC**	**Differential characteristics**
colmated natural cork stopper whose [lenticels are filled] with cork powder	*has_process* [corresponds to LM *'lenticels are filled'*	**ASSOCIATIVE**	colmated natural cork stopper [RESULT] lenticels are filled [PROCESS]	colmated natural cork stopper [SPECIES] = natural cork stopper [GENUS] + filled lenticels [DC]	*/ filled lenticels /*

Sample 7 is the fourth line of Table 26 – Conceptual analysis of Definition 4.

Like Definition 3, this conceptual relation identifier mirrors a semantic dependency of two concepts regarding the associative relation [RESULT-PROCESS]. Once again, the information needed to infer which concept points to a [RESULT] was processed in two stages.

In the first stage, the information was inferred from the meaning of the corresponding LM, since the underlying action, namely "lenticels are filled", connects the meaning of <Colmated natural cork stopper> and <Cork powder>, in a dependency of recipient and process-with-substance, respectively. This allows us to represent the information as:

[colmated natural cork stopper] RESULT *has_process* [lenticels are filled] PROCESS

The transcription into the Aristotelian formula X = Y + DC requires the additional interpretation:

Considering that X $_{SPECIES}$ = Y $_{GENUS}$ + DC, in which the genus = stopper (according to the first line of Table 26) we assume that $_{SPECIES}$ = colmated natural stopper, thus

colmated natural stopper [SPIECES] = stopper [GENUS] + filled lenticels [DC]

Finally, the next interpretation is also centred on the conceptual relation identifier *has_process*; however, based on the Lexical marker "results from", as systematised below in Sample 8. The LM "results from" is a construct from the information inferred from the linguistic analysis and the meaning pointed at by this LM, which is the end of a PROCESS.

Sample 8: Conceptual relation marker has_process and characteristic /dimensional finishing process/

Analysis	Conceptual relation identifier	Conceptual relation	Interpretation	Transcription in X=Y+DC	Differential characteristics
cork powder [results from] the dimensional finishing processes of natural cork stoppers	*has_process* [corresponds to LM *'results from'*]	**ASSOCIATIVE**	cork powder [RESULT] dimensional finishing processes [PROCESS]	cork powder [SPECIES] = natural cork [GENUS] + dimensional finishing process [DC]	*/dimensional finishing process/*

Sample 8 corresponds to the last line of Table 26.

Once again, the conceptual relation identifier *has_process* intends to mirror the semantic dependency between two concepts in an associative dependency relation represented by the dichotomy [RESULT-PROCESS]. As directly made explicit in "dimensional finishing processes", there is an underlying meaning of PROCESS. On the other hand, the meaning of [RESULT] is assigned by the meaning of end-of-a-process conveyed by the LM "results from" to the concept <Cork powder>; contrasting with the means-for-the-result assigned to "dimensional finishing processes". This interpretation entails the following representations:

[cork powder] $_{RESULT}$ *has_process* [dimensional finishing processes] $_{PROCESS}$

Considering that the origin of <Cork powder> is <Natural cork stopper> (a piece of information obtained from the linguistic analysis of the connective form "of" between "finishing processes" and "natural cork stoppers"), we assume that a subsumption [GENUS-SPECIES] is in place, namely:

[cork powder] $_{SPECIES}$ *is_a* [natural cork] $_{GENUS}$

From here, we can finally formulate the transcription into X $_{SPECIES}$ = Y $_{GENUS}$ + DC as follows:

knowing that

$_{SPECIES}$ = cork powder, and $_{GENUS}$ = natural cork,

we assume that DC= dimensional finishing process

that is,

cork powder [SPECIES] = natural cork [GENUS] + dimensional finishing process [DC].

From the analysis of Definition 4, we realised how important the concept `<FinishingProcesses>` is in this domain, as well as its subordinated concepts, from which the denoted manufacturing stages are made explicit in the designations of concepts. For instance, and considering the object defined in Definition 4, a `<ColmatedNaturalCorkStopper>` is a `<NaturalCorkStopper>` that was submitted to the operation of `<Colmation>`. It is the addition of the characteristic /colmated/, which is the differential characteristic, that differentiates `<ColmatedNaturalCorkStopper>` from `<NaturalCorkStopper>` in the concept system.

5.5. A brief overview

Table 27 summarises all the conceptual relations we have inferred during the linguistic analysis of the four definitions addressed in this study.

Table 27: Overview of the conceptual relations inferred from lexical markers

Lexical marker	Conceptual relation identifier	Conceptual relation	A typology of definitional texts governed by the DC
'is a'	*is_a*	**SUBSUMPTION**	stopper [SPECIES]= product [GENUS] + **[any DC added to the genus]**
'commonly referred as'	*is_a*	**SUBSUMPTION**	colmated natural stopper [SPECIES] = natural cork stopper [GENUS] + colmated **[DC added to the genus]**
'is a'	*is_a*	**SUBSUMPTION**	colmated natural cork stopper [SPECIES] = stopper [GENUS] + **[any DC added to the genus]**
'intended to'	*has_function*	**ASSOCIATIVE**	stopper [SPECIES] = product [GENUS] + to seal bottles **[FUNCTION=DC]**

'obtained from'	*has_raw_material*	**ASSOCIATIVE**	stopper [SPECIES] = product [GENUS] + natural cork **[SUBSTANCE=DC]**
'obtained from'	*has_raw_material*	**ASSOCIATIVE**	stopper [SPECIES] = product [GENUS] + agglomerated cork **[SUBSTANCE=DC]**
'obtained from'	*has_substance*	**ASSOCIATIVE**	natural cork [SPECIES] = cork [GENUS] + natural **[SUBSTANCE=DC]**
'obtained from'	*has_substance*	**ASSOCIATIVE**	natural cork [SPECIES] = cork [GENUS] + agglomerated **[SUBSTANCE=DC]**
'intended to'	*has_function*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + to seal containers **[FUNCTION=DC]**
'piece of'	*has_substance*	**ASSOCIATIVE**	stopper [SPECIES] = piece [GENUS] + cork **[SUBSTANCE=DC]**
'usually'	*has_shape*	**ASSOCIATION**	stopper [SPECIES] = piece of cork [GENUS] + cylindrical **[SHAPE=DC]**
'usually'	*has_shape*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + conical **[SHAPE=DC]**
'usually'	*has_shape*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + prismatic quadrangular **[SHAPE=DC]**
'sometimes with'	*has_process*	**ASSOCIATION**	stopper [SPECIES] = piece of cork [GENUS] + rounded edges **[PROCESS=DC]**
'sometimes with'	*has_process*	**ASSOCIATIVE**	stopper [SPECIES] = piece of cork [GENUS] + chamfered edges **[PROCESS=DC]**
''consisting entirely of'	*has_substance*	**ASSOCIATIVE**	natural cork stopper [SPECIES] = stopper [GENUS] + natural cork **[SUBSTANCE=DC]**
'consisting entirely of'	*has_substance*	**ASSOCIATIVE**	natural cork [GENUS] = cork [GENUS] + natural **[SUBSTANCE=DC]**
'submitted to'	*has_process*	**ASSOCIATIVE**	? [SPECIES] = natural cork stopper [GENUS] + sealing operation [DC]
'is made of'	*has_raw_material*	**ASSOCIATIVE**	colmated natural cork stopper [SPECIES] = stopper [GENUS] + natural cork **[SUBSTANCE=DC]**
'is made of'	*has_substance*	**ASSOCIATIVE**	colmated natural cork stopper [SPECIES] = natural cork stopper [GENUS] + colmated **[SUBSTANCE=DC]**
'its lenticels are filled'	*has_process*	**ASSOCIATIVE**	colmated natural cork stopper [SPECIES] = natural cork stopper [GENUS] + filled lenticels **[PROCESS=DC]**
'results from'	*has_process*	**ASSOCIATIVE**	cork powder [SPECIES] = natural cork [GENUS] + dimensional finishing process **[PROCESS=DC]**
'consisting of'	*has_part*	**PARTITIVE**	stopper [SPECIES] = product [GENUS] + one piece **[PARTS=DC]**
'obtained from'	*has_part*	**PARTITIVE**	stopper [SPECIES] = product [GENUS] + several pieces **[PARTS=DC]**
'consisting of'	*has_part*	**PARTITIVE**	stopper [SPECIES] = piece of cork [GENUS] + one element **[PARTS=DC]**
'consisting of'	*has_part*	**PARTITIVE**	stopper [SPECIES] = piece of cork [GENUS] + several elements **[PARTS=DC]**

While we were summarising this information, we observed that depending on the type of conceptual relation, DC tend to be headed by the same label (e.g., [PROCESS=DC]), regarding the relations of subsumpion and partitive relations. For the latter, the label is obviously [PARTS] while as for the relation of subsumption, DC can convey an overwhelming amount of information if the definitional text does not follow the model of an intensional definition, as recommended by ISO.

In a given textual definition following the formula of an intensional definition, DC can be any characteristic, depending on the axis of analysis of the definition, i.e., depending on what is added to the intension of the genus in order to understand the place of the concept being defined, in the concept system. For instance, in the second example for subsumption, in Table 27:

'*commonly referred as'*	*is_a*	**SUBSUMPTION**	colmated natural stopper [SPECIES] = natural cork stopper [GENUS] + colmated **[DC added to the genus]**

the DC added to the genus is a property of the substance; however, such information is activated from the knowledge we acquired from our readings of the specialised texts of the domain. But then again, if we did not take into consideration our knowledge of the domain, it would be possible, through the linguistic analysis and subsequent interpretation of texts, to infer the conceptual relation straightforwardly denoted by the lexical marker "is commonly referred as", which is clearly pointing to a specification of the genus. Hence, the observation of a subsumption.

The same happens with the associative relation although it involves several axes of analysis. Here, DC share identical semantic labels but in a more productive diversity given the prolific semantic dependency identified between concepts, namely [SUBSTANCE]; [FUNCTION]; [PROCESS] and [SHAPE]. Nonetheless, all these semantic labels required a previous analysis of the lexical markers and corresponding co-text, followed by an interpretation of the conceptual markers they point at, as demonstrated in the linguistic analysis followed then by the conceptual analysis, of the four definitions we have addressed.

The observation of the recurrent presence of the aforementioned semantic labels has led us to assume that domain-specific intensional definitions should focus on the knowledge conveyed by those labels in order to organise the concepts coherently.

In view of the above, a methodology for the terminological organisation of this domain is proposed in this study, based on the observed semantic labels that are simultaneously the axis of analysis to build the ontology of the domain (Section 6, p. 215). From our perspective, such methodology embodies the double dimension of Terminology, where the linguistic and the conceptual information complement each other without overlapping. However, and yet again, the primary task necessary to grasp the expert's conceptualisations is the interpretation of texts produced in the specialised context of communication, for texts are the privileged channel to convey knowledge. Hence, the relevance of making the linguistic analysis before embracing the conceptual organisation.

Building the ontology

6. Building the ontology

OntoCork is an ontology[134] in which the concepts of the domain of cork are systematised through logic constructs. Given the vast domain of cork, we narrowed the number of concepts to the ones we have demonstrated so far and for which we have analysed the four definitions (see Sections 4 and 5). Nonetheless, beyond those four concepts, an additional concept will be addressed at the end of this section to demonstrate scalable compositionality.

For this task, we used the ontology editor Protégé[135], a "free, open-source ontology editor and framework for building intelligent systems" – a Standford University project that follows the recommendations of OWL 2 Web Ontology Language and RDF[136] specifications from the World Wide Web Consortium. Protégé is widely used by a community of several areas of interest to build knowledge-based solutions in various spheres, such as biomedicine, e-commerce, and organisational modelling, among others.

This ontology editor is one of the final environments we used to organise the knowledge we had captured from both the linguistic and the conceptual analysis of textual definitions. Based on the five axes of analysis we have retained (see Map 0.0.0, Section 5.2.1.), we have elaborated the five core conceptual relations, as follows:

---

[134] An "[o]ntology is a fundamental form of knowledge representation about the real world. In the computer science perspective, ontology defines a set of representational primitives with which to model a domain of knowledge or discourse (Gruber 2008). The representational primitives of the ontology contain classes, attributes (properties) and relationships between classes. They are used to model knowledge of particular application domains. Ontology sometimes is regarded as for conceptual analysis and domain modeling (Guarino 1998). It is used to analyze the meaning of an object in the world, of a particular domain, and provides a formal specification to describe the object. The object is being "conceptualized" in this case. Gruber (1992) provided a very short definition about ontology – "An ontology is a specification of conceptualization". The formal specification is in support of some sort of knowledge representation model, being generated, analyzed, and processed by computer. The conceptualization has been defined in AI researches (Genesereth and Nilsson 1987, Nilsson 1991) as a structure of <D, R>. The structure defines D as a domain and R as a set of relations on the domain D. This suggests that ontology and conceptualization process are created as domain dependent and relational based." (Lim, Liu, & Lee, 2011, pp. 6-7).

[135] https://protege.stanford.edu/.

[136] Resource Description Framework (see W3C, 2014).

Table 28: Five core conceptual relations of the ontology.

Axis of analysis	Format relation in Protégé	Type of conceptual relation
FUNCTION	hasFunction	associative relation, subtype [OBJECT-FUNCTION]
SUBSTANCE	IsMadeOf	associative relation, covering both sub-types [RAW MATERIAL – PRODUCT] and [MATTER/SUBSTANCE – PROPERTY]
PARTS	hasStructure	partitive relation [PART-WHOLE]
FINISHING PROCESS	hasProcess	associative relation, within the sub-type [PROCESS-RESULT]
SHAPE	hasShape	associative relation, sub-type [OBJECT-SHAPE]

As we will see in this section, additional conceptual relations will be taken into consideration, such as a sub-type of the isMadeOf relation (see Table 28).

All conceptual maps elaborated during the conceptual analysis served us as the first draft to model the ontology. The conceptual maps elaborated with CmapTools are non-formal representations and for this, we have resorted to the characteristics that underlie the *criteria of subdivision*[137] for the systematisation of concepts. As we will demonstrate in the next lines, the criterion of subdivision is a methodology also present in the task of building the ontology.

Before we start to describe the method used to build the ontology, we will first address a few graphical representations (e.g., OntoGraf[138]), here considered formal given the underlying logic constructs. Our purpose is to straightforwardly demonstrate how concepts, characteristics and axes of analysis are put in place. The logic constructs

[137] According to (ISO/FDIS 1087), the criteria of subdivision is also known as "subdivision criterion: type of characteristic according to which a superordinate concept is divided into subordinated concepts." (2019 (E), p. 5).

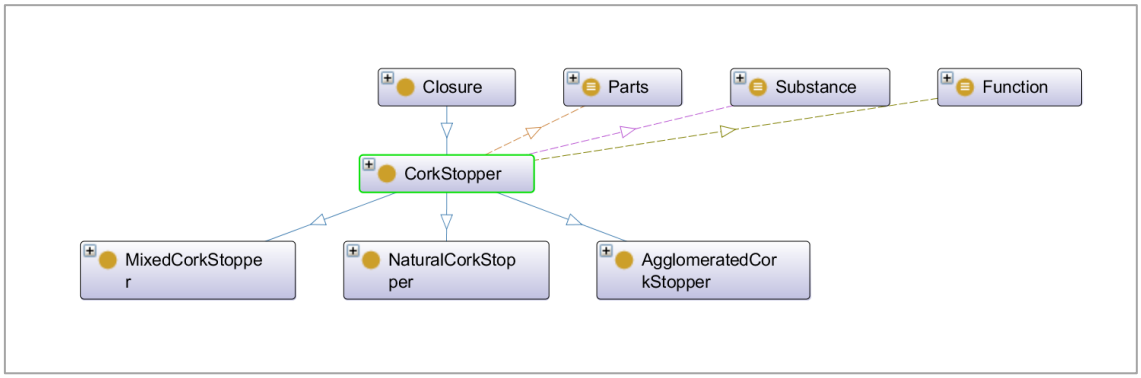
[138] "OntoGraf gives support for interactively navigating the relationships of your OWL ontologies. Various layouts are supported for automatically organizing the structure of your ontology. Different relationships are supported: subclass, individual, domain/range object properties, and equivalence. Relationships and node types can be filtered to help you create the view you desire." (StanfordEdu, 2020).

to define each concept pertaining to the four definitions under analysis will be addressed in Section 6.2 (p. 221).

6.1. From CmapTools to Protégé – Definition 1: <Stopper>

We will start with the axes of analysis captured in Definition 1: <Stopper> (see Section 4.2.1, p. 132).

The following representation is an OntoGraf, a plug-in of the Protégé editor.



OntoGraf 1: Ontological representation of <CorkStopper>, a concept that can be defined according to the 3 axes of analysis inferred from Definition 1. The relation of subsumption (genus-differentia) is vertically represented with blue coloured full-line arcs, while for other conceptual relations, arcs are dashed. Their different colours represent different types of relations, e.g., the pink dashed line between CorkStopper and Substance represents the owl:ObjectProperty IsMadeOf, which corresponds to the associative relation sub-type [MATTER/SUBSTANCE-PROPERTY] in our study.

OntoGraf allows users to visualise the ontology. This plug-in does not show the reasoner[139] inference. However, we considered it useful to demonstrate our work regarding hierarchical, partitive and associative conceptual relations.

OntoGraf 1 corresponds to the ontological representation of Conceptual Map 1 (see Section 5.1.1.) according to the "Types of cork stopper".

Moving away from the conventions used in CmapTools, in Protégé instead we write the name of the concepts without the underscore, e.g., `NaturalCorkStopper`. Furthermore, and similarly to the decision made in CmapTools, we decided to name the

[139] also called *classifiers* – thus, the term *classification* when a concept is inferred as pertaining to a given class of concepts.

genus concept `CorkStopper`, instead of `Stopper`. In our opinion, the location of the concept `CorkStopper` in the concept system is less ambiguous than `Stopper`, for the latter seems as generic as the superordinate concept `Closure`. Furthermore, given our awareness of the existence of other kinds of stoppers as to the substance they are made of, e.g., plastic stopper, we decided to make explicit the characteristic /cork/ within the name of the concept.

As depicted in OntoGraf 1, the concept `CorkStopper` is modelled as a specialisation (s*pecies*) of the concept `Closure`, the closest generic concept (*genus proximum*), thus having a relation of *subsumption* with `Closure`. This relation is also known as *is-a* relation or *genus-differentia*[140] relation. `CorkStopper` is simultaneously the *genus* of three sub-concepts, namely `MixedCorkStopper`, `AgglomeratedCorkStopper` and `NaturalCorkStopper`, meaning that the latter are subsumed by the former therefore co-relating as siblings – also known as coordinated concepts. This systematisation stems from the three types of `Substance` that `CorkStopper` relates with, by virtue of the differential characteristics /natural cork/, /mixed cork/ and /agglomerated cork/.

`CorkStopper` also entertains conceptual relations[141] with `Parts`, `Substance` and `Function`, namely through a partitive relation with the first, and an associative relation with the last two. These relationships underpin the logic constructs to model the information: a `CorkStopper` can be formally defined according to its constituent `Parts`, the `Substance` it is made of, and the `Function` it has.

`Parts`, `Substance`, and `Function` represent three of the five axes of analysis we have retained as mentioned above. We have chosen to insert these three axes in the

[140] According to Smart (1849, p. 146), "All knowledge consists in being aware of the relations in which the thing known stands to oneself and to other things; which are said to be of the same genus or kind; and then to distinguish it from these things by stating its difference. Genus and difference [...] form a definition." This *genus-differentia* relation is therefore of utmost importance in our study, not only for the organisation of knowledge but also for the analysis and/or writing of definitions, either in formal (conceptual) or informal (natural language) format.

[141] In our study, "conceptual relations" correspond to owl:ObjectProperty, in OWL-DL – the formal language used in Protégé. According to W3C, "A property is a binary relation. Two types of properties are distinguished: (1) *datatype properties*, relations between instances of classes and RDF literals and XML Schema datatypes; (2) *object properties*, relations between instances of two classes." (W3C, 2004).

ontology as generic concepts, each described by enumeration[142], i.e., through "*an exhaustive enumeration of individuals that together form the instances of a class"* (W3C, 2004). In our study, instead of classes, we will refer to concepts, and the extension[143] of each of these concepts is represented in curly brackets, as follows:

(1) `Parts` = {Body, Disc};

(2) `Substance` = {NaturalCork, AgglomeratedCork};

(3) `Function` = {ForStillWines, ForSparklingWines}.

The following schema, in OWL-XML[144] – an excerpt of the OWL file of the ontology – is the description of `Function`, by enumeration:

```
Class IRI="#Function"/
ObjectOneOf
    NamedIndividual IRI="#ForSparklingWines"/
    NamedIndividual IRI="#ForStillWines"/
/ObjectOneOf
```

As we can see above, this option describes the concept `Function` by listing all the members pertaining to the extension of this concept, in this case, mirroring all the functions that cork stoppers are manufactured for (i.e., for different types of wine). Using this type of concept enumeration, we can formally assert *necessary and sufficient conditions*[145] for class membership, in this case, consisting of an enumeration of two individuals, no less, no more (see W3C, 2020). For this study, we will not enumerate

[142] According to W3C, "Classes can be described by enumeration of the individuals that make up the class. The members of the class are exactly the set of enumerated individuals." (W3C, 2004).

[143] In the sense of Description Logics : "Les entités de base qui sont définies et manipulées dans une logique de descriptions sont les concepts et les rôles. Un concept dénote un ensemble d'individus - l'extension du concept - et un rôle dénote une relation binaire entre individus. Un concept possède une description structurée qui se construit à l'aide d'un ensemble de constructeurs introduisant les rôles associés au concept et les restrictions attachées à ces rôles." (Napoli, 1997, p. 8).

[144] For more details, see https://www.w3.org/TR/owl-xmlsyntax/

[145] Our decision is grounded on the notion that "It is […] possible to make the distinction [between defined and partially defined] using the difference between "SubClassOf" and "EquivalentClasses" (Rector, et al., 2004). In Protégé, a "defined class" corresponds to a concept with a "complete" definition – membership requires necessary & sufficient conditions, while a "primitive class" corresponds to a concept with a partial definition – membership requires only necessary conditions. "It is critical to understand that, in general, nothing will be inferred to be subsumed under a primitive class by the classifier" (*ibid.*).

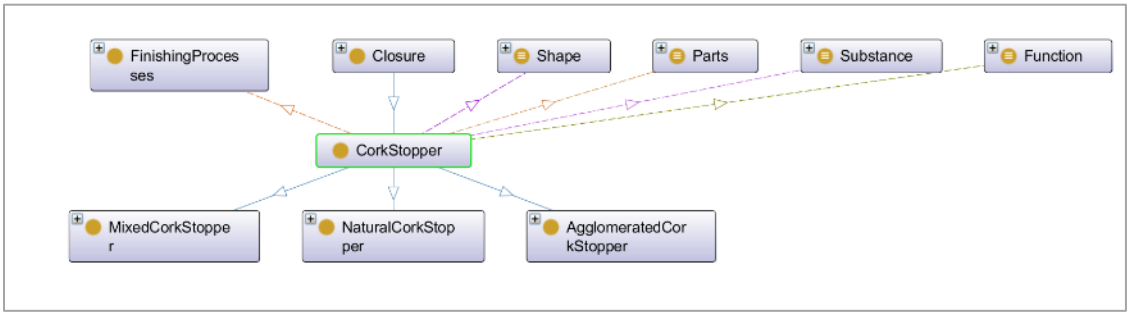
more than two types of functions because they are enough to describe the typology of cork stoppers under analysis.

The relevance of necessary and sufficient conditions is that

> necessary characteristics hold for all objects in the extension of a concept, i.e. they correspond to properties that all objects in the extension must have [while a] sufficient characteristic is one of a set of characteristics that determines whether a specific object belongs in the extension of a given concept. A sufficient characteristic is not necessarily true of all objects in the extension of the concept, but any object having the properties corresponding to the characteristics in this set belong to the extension of the concept. (ISO 704, 2009, p. 15)

In light of what we have said above, by *necessary and sufficient conditions*, we mean that the definition becomes complete – the list of all the conditions required for membership – as opposed to being a partial definition – where at least one of the necessary conditions required for membership in the sense of Description Logics (see Rector, et al., 2004).

In the next OntoGraf, we can see that we have finally assembled the 5 axes of analysis obtained from both Definition 1 and Definition 2 – two definitions that complement the information regarding <Stopper>.



OntoGraf 2: Ontological representation of concepts holding five relations according to the 5 axes of analysis drawn from the conceptual analysis of Definition 1.

OntoGraf 2 is a representation in Protégé of all the associative conceptual relations we have drawn from the conceptual analysis, including subsumption (*genus-differentia*) and partitive, which were already included in OntoGraf 1.

Based on Conceptual Map 2 (Section 5.2.1., p. 194), and following the evolution of CMap 0.0.0 (Section 5.2.1., p. 191), two concepts were added to the ontology: (1)

FinishingProcesses and (2) Shape to account for the different formats that a CorkStopper can have. Thus, a CorkStopper can be defined according to its Parts, Substance, Function, FinishingProcesses and Shape. These five concepts are what we refer to as axes of analysis, and for that reason, we chose to describe most of them by enumeration. The last concept described by enumeration is

Shape = {Chamfered, Conical, Cylindrical, PrismaticQuadrangular, Round}

The axis of analysis FinishingProcesses was treated differently. This concept subsumes 21 sub-concepts: 2 are direct sub-concepts that, in turn, subsume 3 and 4 sub-concepts, and so forth. The option of not describing this concept by enumeration is tied with the complex axiom[146] constructs we are aiming at to describe related concepts. CorkStopper relates to FinishingProcesses in a wide intricacy with the associative relation hasFinishingProcess and corresponding sub-types. This intricacy allows us to construct complex concept descriptions and to reason with Protégé. The topic of complex constructs will be addressed in Section 6.3 (p. 237).

6.2. The formal description and annotations of CorkStopper in Protégé

In the following lines, we will address to the logic constructs we have construed to formally describe in Protégé the concepts we have analysed in Definitions 1, 2, 3, and 4 (Sections 4 and 5).

Following the same order of the concepts defined by the above four definitions, we will first demonstrate the logical constructs for CorkStopper, which is the most generic concept within the <Cork stopper> typology right after <Closure>.

[146] According to W3C, "Axioms are used to associate class and property identifiers with either partial or complete specifications of their characteristics, and to give other information about classes and properties". Information available online at (W3C, 2004).

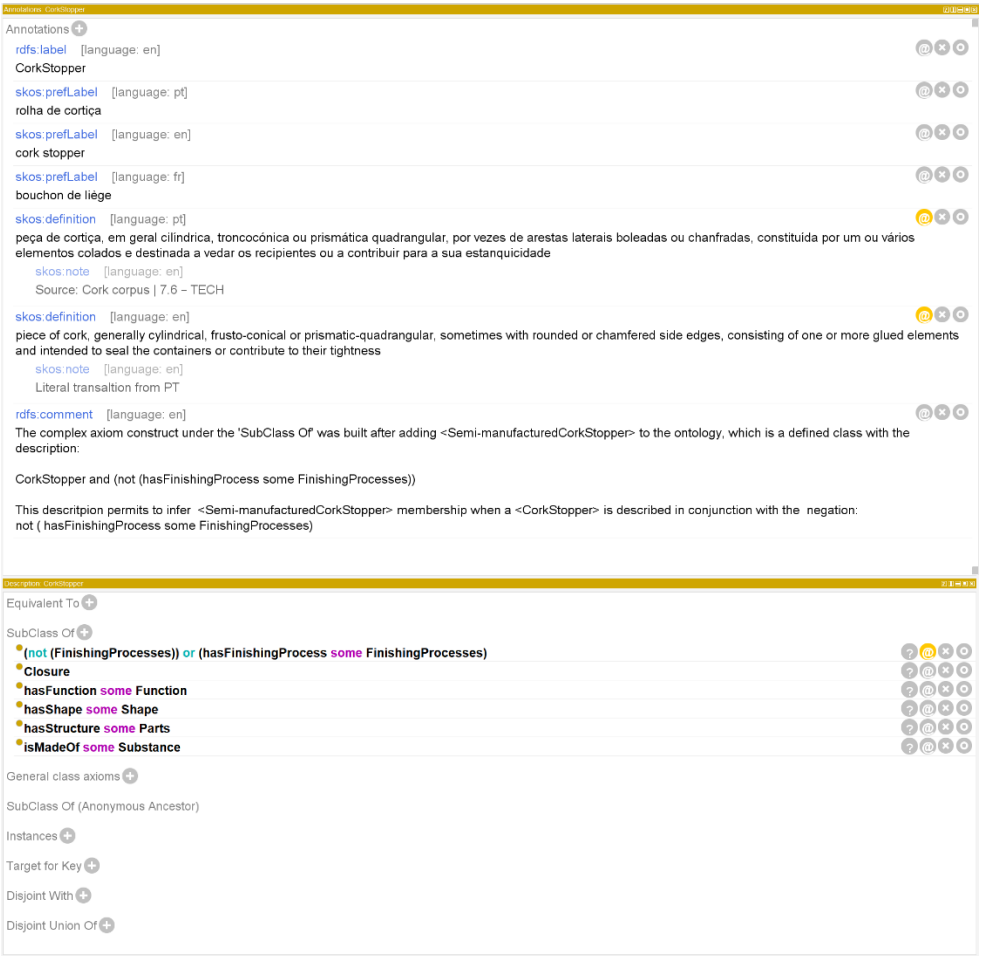


Figure 37: Annotations (top) and axiom constructs (bottom) of the characteristics of the concept CorkStopper in Protégé.

In Figure 37, we represent the visualisation of the concept annotations[147] (top panel) and the concept editor (bottom panel) in Protégé. In the latter, we can see axioms constructs describing the characteristics of the concept `CorkStopper`, which we have retained for this generic concept and correspond to the 5 axes of analysis. On the top panel, we can see the definition of the concept `CorkStopper` written in natural language, both in PT and EN, as well as the verbal designation of the concept, namely,

[147] Annotations are inserted as labels in a (machine) interoperable format: "One common set of additional tags that [are] included here are some of the standard Dublin Core metadata tags. The subset includes those that take simple types or strings as values. Examples include Title, Creator, Description, Publisher, and Date." (W3C, 2003).

the term in PT, and the equivalents in EN and FR. For the edition of their metalanguage, we have used the SKOS Core Vocabulary, e.g., skos:definition; skos:prefLabel[148] .

As shown at the bottom of Figure 37, the axes of analysis of `CorkStopper` are described through axioms and restriction constructs[149] resorting to OWL-DL Boolean operators, under SubClass Of. Here, `CorkStopper` is defined as a primitive concept[150], which is subsumed by `Closure` and holds partitive and associative relations (owl:ObjectProperties) with `Function`, `Shape`, `Parts`, `Substance` and `FinishingProcesses` – the 5 main axes of analysis.

To express those 5 axes of analysis with restriction constructs in OWL-DL, we have used the relations hasStructure, hasFunction, hasShape, hasFinishingProcess and isMadeOf (owl:ObjectProperties, according to OWL terminology). These relations correspond to partitive, in the first case, and associative types, in the other cases, and they all play a restricting role for the systematisation of concepts. The role of these restrictions is to formally describe, in the sense of OWL-DL, which characteristics a concept must comprise to be considered a specialisation (a kind of) of a given *genus* concept. The following schema aims to demonstrate what we have just described.

[148] "The SKOS Core Vocabulary is an application of the Resource Description Framework (RDF). RDF provides a simple data formalism for talking about things, their properties, inter-relationships, and categories (classes). Using RDF allows data to be linked to and/or merged with other RDF data by Semantic Web applications." (W3C, 2005).

[149] According to Rector, "restrictions [are] constructed as quantified role-concept pairs, e.g. (restriction hasLocation someValuesFrom Leg) meaning "located in some leg". (2003, p. 2).

[150] Primitive concepts are described by necessary conditions and are at the basis of the construction of defined concepts, i.e., those that have a definition (see Rector A. L., 2003; Napoli, 1997).

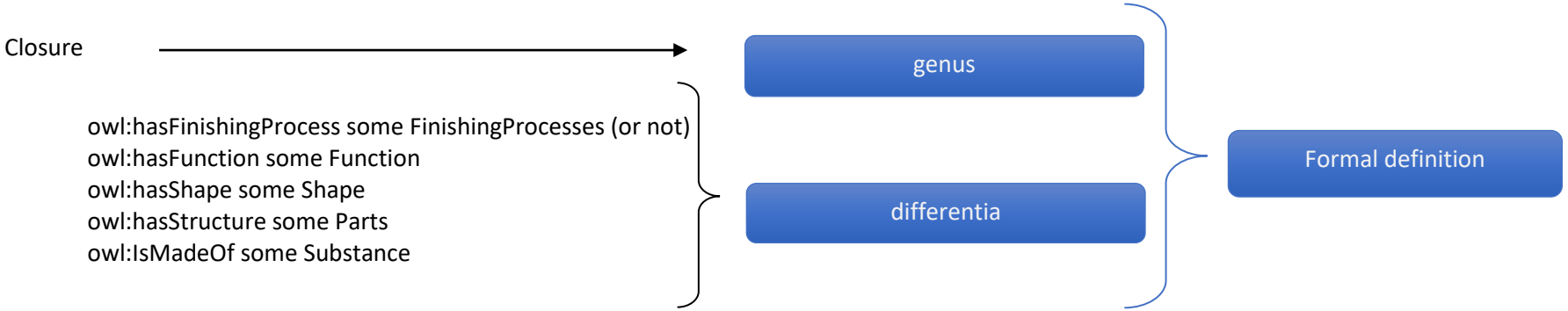


Figure 38: Characteristics of CorkStopper - a specialisation (a kind) of Closure.

Figure 38 above is a manually built reproduction of what we had previously seen in Figure 37 – the class editor in Protégé – for the description of the concept `CorkStopper`. The purpose of this schema is to observe how formal definitions are built in OWL-DL, resorting to this ontology editor. In this example, `Closure` is the genus and the block of relations (owl:ObjectProperties) are the differential characteristics that allow us to define this specific kind of closure, i.e., the `CorkStopper`. We can observe here again, the Aristotelian formula, in which X [cork stopper] = Y [closure] + DC [function; shape; parts; substance; finishing process (or not)]. The last DC is a complex construct: "hasFinishingProcess some FinishingProcesses (or not)" and will be discussed in Section 6.3.1. (p. 239).

For the explanation of the formal constructs in “Manchester OWL syntax“ (W3C, 2012) shown above in Figure 38, we will choose the description of the three concepts that a `NaturalCorkStopper` subsumes according to the `Parts` combination. This means that the subdivision criteria are the characteristics /one piece/ or /several pieces/, albeit differently asserted as we will demonstrate.

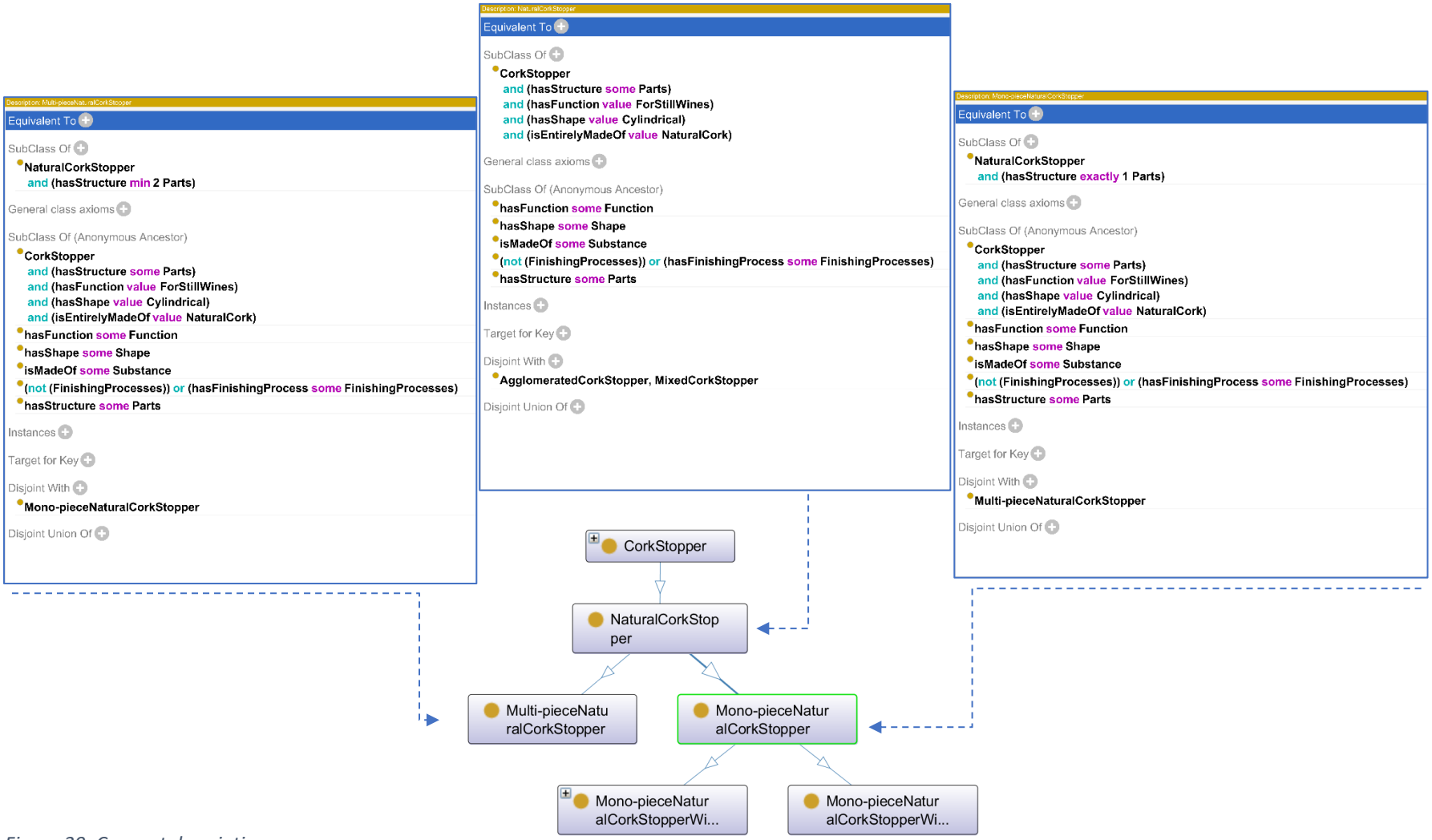


Figure 39: Concept description

of <NaturalCorkStopper>, <Mono-pieceNaturalCorkStopper> and <Mono-pieceNaturalCorkStopper> along with their corresponding description in Protégé.

We have assembled above in Figure 39 two types of concept representation, namely an ontological representation of the three concepts `NaturalCorkStopper`, `MonoPieceNaturalCorkStopper`, and `MultiPieceNaturalCorkStopper` as an OntoGraf, along with their corresponding descriptions[151] in the concept editor.

As depicted above, `NaturalCorkStopper` is a specification of `CorkStopper`, the *proximum genus;* therefore, the former inherits the latter's characteristics, as we can observe on the concept description editor (top centre class description view), under the heading "SubClass Of (Anonymous Ancestor)". This means that the description of `NaturalCorkStopper` is more specific – as shown under "Subclass Of" – in which its *proximum genus* `CorkStopper` is made explicit along with additional characteristics. These additional characteristics are asserted through logical constructs, as explained in the following lines.

6.2.1. The description of <NaturalCorkStopper> in Protégé

We will now explain the constructs we have asserted to describe `NaturalCorkStopper` by the same order we can observe them in the next concept description, in Figure 40 (one of the three descriptions, extracted from Figure 39 above).

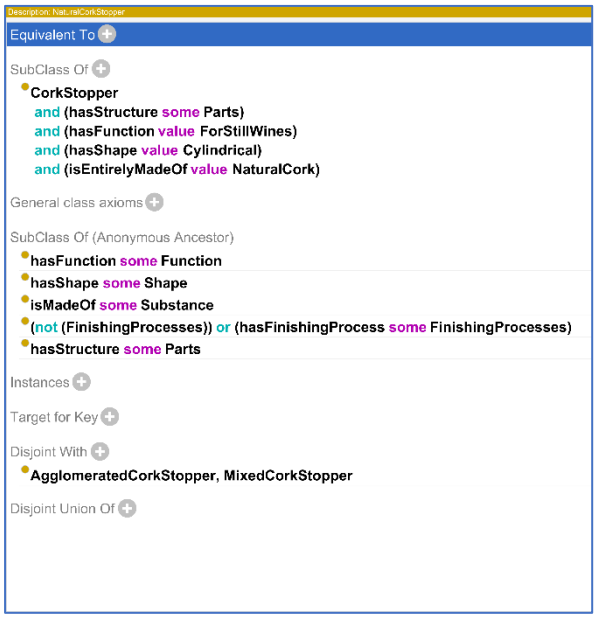


Figure 40: Description of <NaturalCorkStopper> in the class editor of Protégé.

[151] Named "class description view" in Protégé terminology: "The class description view is the core of the class editor. It allows the logical description of the selected class to be edited using Manchester OWL Syntax". Source available online (Protégé, 2020).

For the description of `NaturalCorkStopper`, we decided to assert that there are two different kinds of such concept, depending on the structure they are composed of, i.e., the number of their compositional parts. To logically assert this information, we wrote the first axiom, in Manchester OWL Syntax, as the following construct:

(1) (hasStructure some Parts)

In axiom construct (1), the operator some[152] allows the description of those concepts that are composed of at least 1 part or more. Hence, the interpretation of axiom construct (1) is: `NaturalCorkStopper` is structurally composed of at least one `Part` or several `Part`, by means of the partitive relation (owl:ObjectProperty) hasStructure.

The remaining characteristics are logically asserted through the following constructs:

(2) hasFunction value ForStillWines

(3) hasShape value Cylindrical

(4) isEntirelyMadeOf value NaturalCork

The constructs (2), (3) and (4) are simple axioms and attend three axes of analysis of the concept description via subsumption. This hierarchical relation is expressed through the intersection of two concepts. For instance, the *proximum genus* `Closure` intersects `ForStillWines` by means of the associative relation hasFunction and the local value restriction of the property owl:hasValue[153], viz. value. This last owl property means "one of"[154] and is used to constrain the associative relation, as below enumerated for each axiom:

---

[152] The operator *some*, stands for "Existential Restriction", a constructor that is also expressed as owl:someValuesFrom: an OWL value constraint that restricts the range of a property when used with a specified class" (Lacy, 2005, p. 186).

[153] According to Lacy, "the 'owl:hasValue' property can be used to define classes based on the property values of its individual members. At least one of the individual's property values must be equal to the individual or data value identified by the 'owl:hasValue' constraint." (2005, p. 232).

[154] Linguistic expression in natural language based on the property "owl:oneOf" and according to: "The value of the "owl:oneOf" property is a list of individuals that exhaustively identifies the class extension. The [enumeration] is used to specify the closed list of individuals." (Lacy, 2005, p. 225).

(2.1) While the *genus* (the anonymous ancestor – i.e., `CorkStopper`) is generically described by the construct hasFunction some `Function`, its *species* `NaturalCorkStopper` is described by that same relation yet associated with a specific concept by means of a value restriction. This restriction is used, in this context, to constrain the associative relation hasFunction to only one of the concepts pertaining to the extension[155] of `Function`; which is exactly that of `ForStillWines`.

(3.1) While the *genus* (the anonymous ancestor) is generically described with the relation hasShape some `Shape`, its *species* `NaturalCorkStopper` is described with the same relation yet associated with a specific concept by means of a value restriction. In this context, the restriction is used to constrain the associative relation hasShape to only one of the concepts pertaining to the extension of `Shape`; which is exactly that of `Cylindrical`[156].

(4.1) While the *genus* (the anonymous ancestor) is generically described by the relation isMadeOf some `Substance`, its *species* `NaturalCorkStopper` is described by the same relation although by means of a value restriction. In this last context, the restriction is used to constrain the associative relation isEntirelyMadeOf to only one of the concepts pertaining to the extension of `Substance`; which is exactly that of `NaturalCork`.

The associative relation isEntirelyMadeOf (owl:ObjectProperty) is a sub-type of isMadeOf (owl:ObjectProperty). Choosing to create this sub-type of associative relation has to do with other kinds of `CorkStopper` under analysis, such as `MixedCorkStopper` as we will further demonstrate.

The concept description of `NaturalCorkStopper` is comprised thus by four axioms. These axioms correspond to what we consider *differential* characteristics, to the extent to which they are what differentiates one concept from another, thus, the systematisation of a new concept is verified. This conceptual systematisation encompasses an intricacy of several types of conceptual relations simultaneously, which

[155] Listed by enumeration in the sense of OWL.

[156] For the sake of simplicity, we shall not describe any other shape apart from cylindrical.

can be hierarchical or not. Regarding that intricacy, Sager refers to what is needed to model knowledge:

> The model [of knowledge] is conceived as a multidimensional space in which intersecting axes represent some kind of conceptual primitives or characteristics. They may also be seen as features or components. A concept, i.e., a unit of knowledge, can be represented and identified uniquely by references to its coordinates along each axis. Listing the values of a concept with respect to each axis, component or feature is equivalent to defining its position in the knowledge space. (1990, p. 15)

In sum, and following the notion of "conceptual primitives or characteristics" necessary to define a concept with regards to its place in the knowledge system – which we rather designate as "concept system" (Wüster, 1998) – the concept of `NaturalCorkStopper` is defined by virtue of the intersection of the four axes of analysis asserted to define its genus and differentiated by the specific values of the axes it relates with. The specific values are the characteristics, which, as already mentioned, underlie the typology of conceptual relations.

The following schema in Figure 41, within which the concept description of `MonoPieceNaturalCorkStopper` is asserted in Protégé, aims at demonstrating this notion of modelling knowledge:

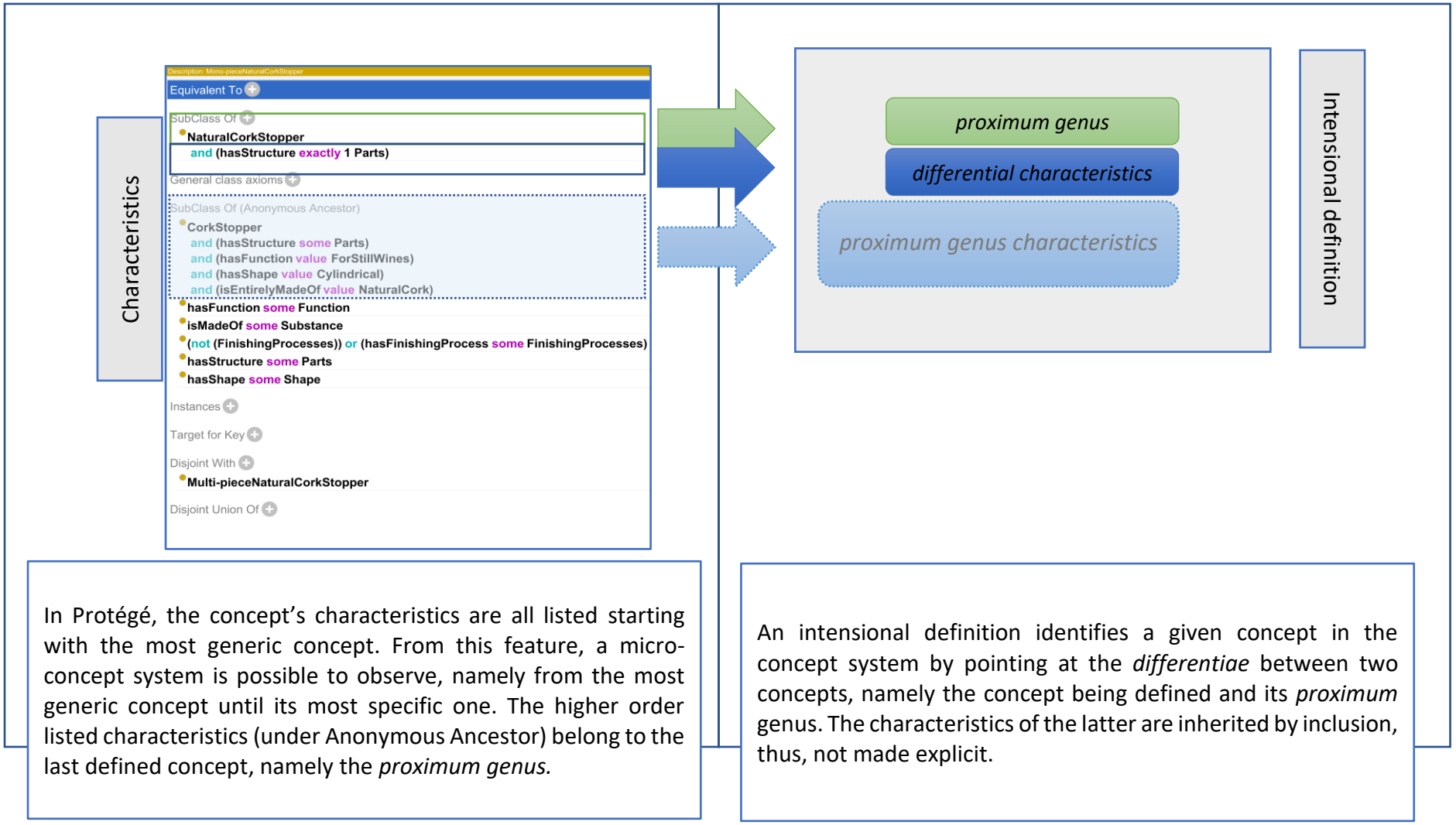


Figure 41: Schema to demonstrate the model of knowledge advocated by Sager, where the characteristics of a given concept are listed according to their axes of references.

6.2.2. The description of <MonoPieceNaturalCorkStopper> in Protégé

As shown below in Figure 42, `MonoPieceNaturalCorkStopper` is one of the *species* of `NaturalCorkStopper`. This concept description was previously shown in Figure 39, Section 6.2 (p. 226).

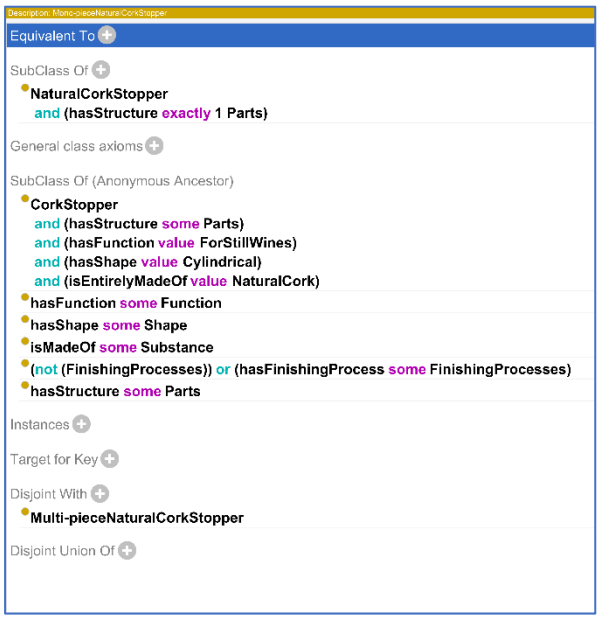


Figure 42: Description of <Mono-pieceNaturalCorkStopper> in the class editor of Protégé.

For the description of `MonoPieceNaturalCorkStopper` we have construed the following axiom:

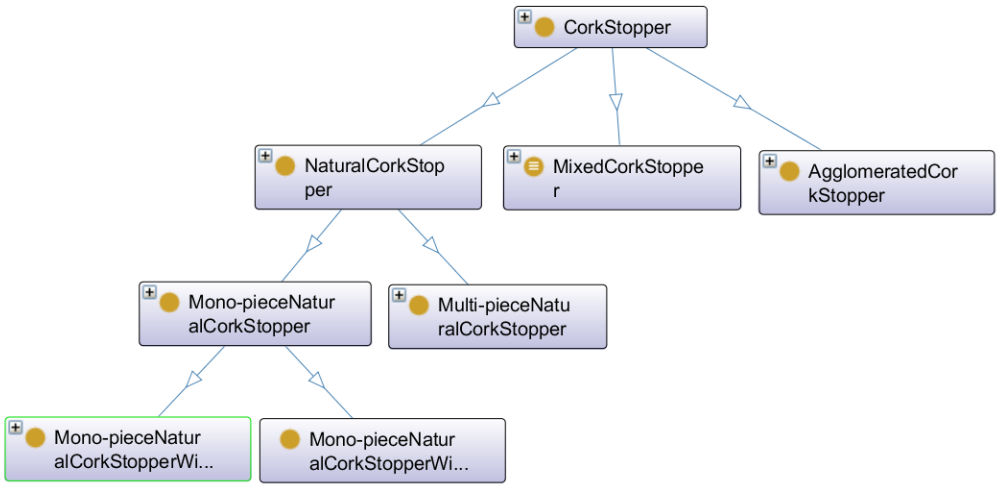
(5) NaturalCorkStopper and (hasStructure exactly 1 Parts)

In Manchester OWL Syntax, the Bolean operator and stands for the property of intersection (owl:intersectionOf). The purpose of using this operator aims at the property of intersection of multiple concepts. This intersection can define a new concept considering that the application of an intersection is analogous to a *logical conjunction.* In other words, an arbitrary number of concepts can be identified since the intersection of concepts includes members that belong to both concepts. Thus, an axiom construct using the property of intersection describes the inclusion of all concepts that are common to the stated concepts (see Lacy, 2005).

In this line of thought, the axiom construct (5) describes the specification of `MonoPieceNaturalCorkStopper` through the intersection of the *genus*

`NaturalCorkStopper` along the partitive relation hasStructure, which restrains the number of `Parts`, in this case, no more than 1 by means of the owl property of cardinality *exactly* (owl:cardinality). In this context, the number of parts corresponds to the differential characteristic from the *genus,* while the remaining characteristics are inherited from the genus, as depicted above in Figure 42.

As previously seen in Conceptual Map 4: <Colmated_cork_ stopper> (Section 5.3, p. 199), the characteristic **/with sealing operation/** and its counterpart **/without sealing operation/** were added by differential division straightforwardly after the *proximum genus*. However, in Protégé, we had to create two interjacent concepts to represent the dichotomy **/with sealing operation/** and **/without sealing operation/**. This necessity results from the existence of kinds of `CorkStopper` without `FinishingProcesses` and corresponding classification, as we will further discuss in Section 6.3.3. (p. 243). Thus, two concepts were named for that purpose, namely `MonoPieceNaturalCorkStopperWithFinishingProcess` and `MonoPieceNaturalCorkStopperWithoutFinishingProcess`: two species of `MonoPieceNaturalCorkStopper` as shown in the next OntoGraf:

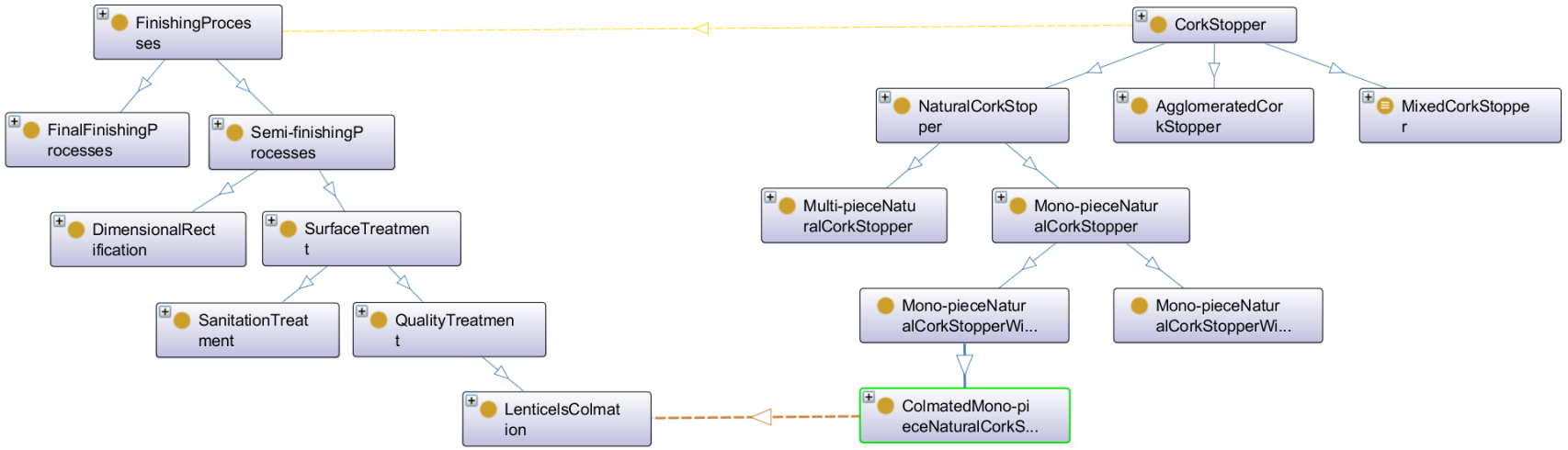


OntoGraf 3: Two interjacent concepts to represent the dichotomy "with or without finishing process" in order to model species of <Mono-pieceNaturalCorkStopper> accordingly.

Once the interjacent concepts were created to represent the dichotomy **/with sealing operation/** and **/without sealing operation/**, we were finally able to formally describe

the concept `ColmatedMonoPieceNaturalCorkStopper` in Protégé. To do so, we had to create an associative relation to convey the characteristic /with sealing operation/, which we named hasLenticelsColmationOperation.

The next representation is the ontological representation of `ColmatedMonoPieceNaturalCorkStopper`, where we can observe several concepts systematised, either vertically – in a hierarchal dependency – or horizontally – in a pragmatic (associative) dependency – according to the differential characteristics.



OntoGraf 4: Ontological representation of <ColmatedMono-pieceNaturalCorkStopper>, a specification of <Mono-pieceNaturalCorkStopper> that was submitted to <LenticelsColmation>, a kind of <FinishingProcesses>. The relation of subsumption is represented with vertical blue arcs and the associative relation, sub-type [PROCESS-RESULT] is represented with horizontal dashed lines. <ColmatedMono-pieceNaturalCorkStopper> and <LenticelsColmation> are linked by the associative relation owl:hasLenticelsColmationOperation, represented in a brown dashed line at the bottom of both hierarchical representations.

As we can see in OntoGraf 4, the ColmatedMonoPieceNaturalCorkStopper is a specification of MonoPieceNaturalCorkStopperWithFinishingProcess. The relation of subsumption is represented with vertical blue arcs and the associative relation is represented with horizontal dashed lines.

Furthermore, ColmatedMonoPieceNaturalCorkStopper and LenticelsColmation are linked by the associative relation, sub-type [PROCESS-RESULT]: hasLenticelsColmationOperation. As mentioned above, this conceptual relation is based on the differential characteristic /with sealing operation/, which was drawn from the analysis of Definition 3, and subsequently used in Conceptual Map 4.

Hence, hasLenticelsColmationOperation is the associative relation that induces the specification of MonoPieceNaturalCorkStopperWithFinishingProcess by *differentia* as shown below in the concept editor:

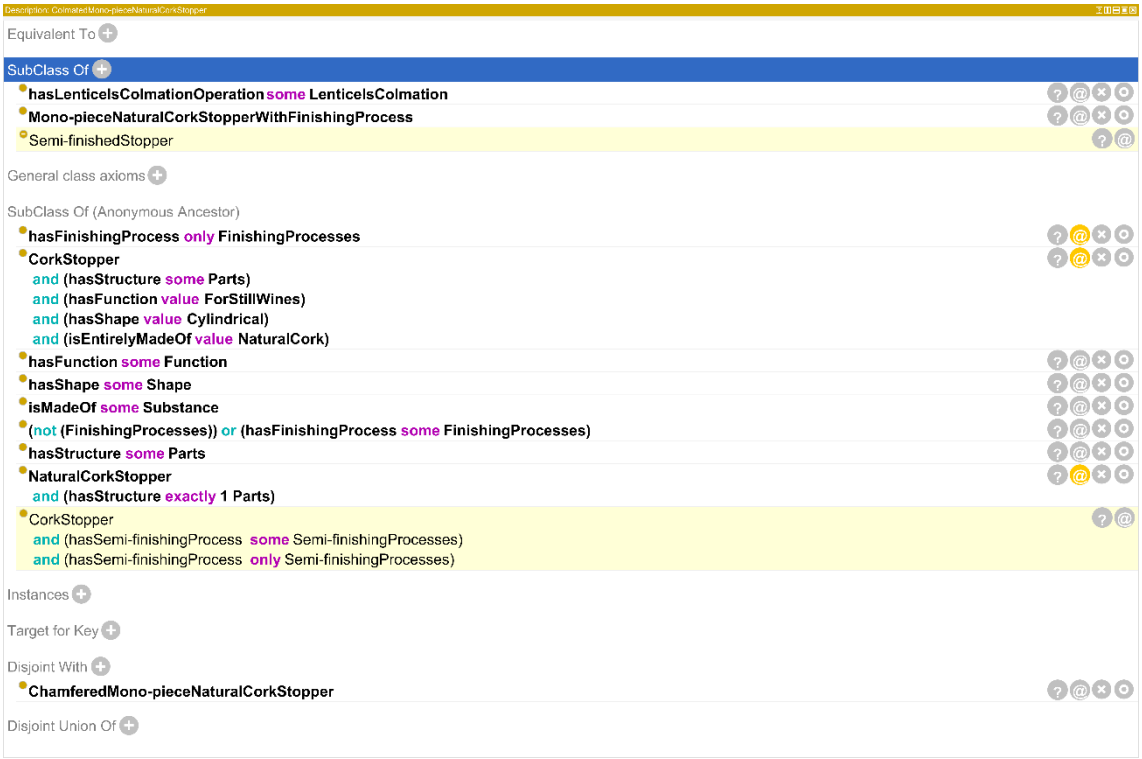


Figure 43: Concept description of <ColmatedMono-pieceNaturalCorkStopper> in Protégé.

Figure 43 depicts the description of the concept ColmatedMonoPieceNaturalCorkStopper, in which we can see the differential characteristic *LenticelsColmation* declared through a simple axiom construct, as represented below:

(6) hasLenticelsColmationOperation some LenticelsColmation.

Axiom (6), in conjunction (or intersection) with its *proximum genus* `MonoPieceNaturalCorkStopperWithFinishingProcess`, is what underlies the description of this new concept: `ColmatedMonoPieceNaturalCorkStopper`, along with the associative relation hasLenticelsColmationOperation. It must be noted that the intersection occurs by default between simple axiom constructs asserted under SubClass Of in the concept description editor, e.g.,

SubClass Of:
MonoPieceNaturalCorkStopperWithFinishingProcess
hasLenticelsColmationOperation some InkMarkingOperation ⟷ "and"

Finally, it is also possible to see above in OntoGraf 4 (p. 235), a hierarchical representation of `FinishingProcesses`, in which the involved operation of the concept we have just described is allocated as the most specific concept of this hierarchy. The interpretation of this subsumption is: `LenticelsColmation` is a kind of `QualityTreatment`, which is a kind of `SurfaceTreatement`, which in turn is a kind of `Semi-finishingProcess`, all of these are kinds of `FinishingProcesses`.

According to the rules we have created to build this ontology in Protégé, such as axiom (6), concepts are classified by the reasoner with regards to their stage of completion, in the manufacturing process. An example of such classification may be seen above in Figure 43: `ColmatedMonoPieceNaturalCorkStopper` is classified by the reasoner as a `Semi-finishedStopper`, highlighted in yellow. The rules and further examples of this type of classification will be addressed in the next section.

6.3. Finishing processes

As remarked in this study, cork stoppers may undergo finishing processes or not. However, another piece of information needs to be added: it all depends on the manufacturing stage of the stopper; therefore, a classification of completion is attributed to the stopper according to the last operation it was submitted to.

A few words must, therefore, be introduced before the demonstration of the systematisation of `FinishingProcesses` in the ontology.

Until it has achieved the state of **finished**, a cork stopper is designated as **semi-manufactured** if it was not submitted to any kind of *finishing treatment*; or **semi-finished**, if not submitted to any kind of *final treatment*. In other words, a cork stopper must be submitted to at least one *final treatment* to acquire the state of *finished* after it had previously been submitted to *semi-finishing treatments*.

In this line of the *finishing process*, and regardless of the type of cork it is made of, or the number of its compositional parts, a cork stopper undergoes several operations until it is a finished product. Cork stoppers may be sold with a semi-finished or finished status. The client acquires them (a winery, for instance) either unready or ready to be used, depending on the client's purposes or means to finish the stoppers. In brief, a semi-finished stopper is a stopper that was submitted to any finishing treatment of the finishing process, such as "rectifying"[157], "washing"[158], and subsequently "drying"[159], except any kind of "final treatment"[160]. At this point, the unready-for-use-stopper is either sold, packed and transported or continues through the finishing process, until it is ready to be used. To be considered a finished product, the stopper must undergo the final treatments, which are branding and/or surface coating treatment.

Finally, it is essential to clarify the origin of our reference to several concepts falling under the classification of `FinishingProcesses`. Given the frequent absence or partial mention of "operations" or "finishing process treatment" in all of the natural language definitions we have drawn from the corpus regarding a cork stopper's typology, we had to search for contextual definitions in the corpus to obtain information regarding these activities.

---

[157] *Rectificação*, in Portuguese.

[158] *Lavagem* or *Lavação,* in Portuguese.

[159] *Secagem,* in Portuguese.

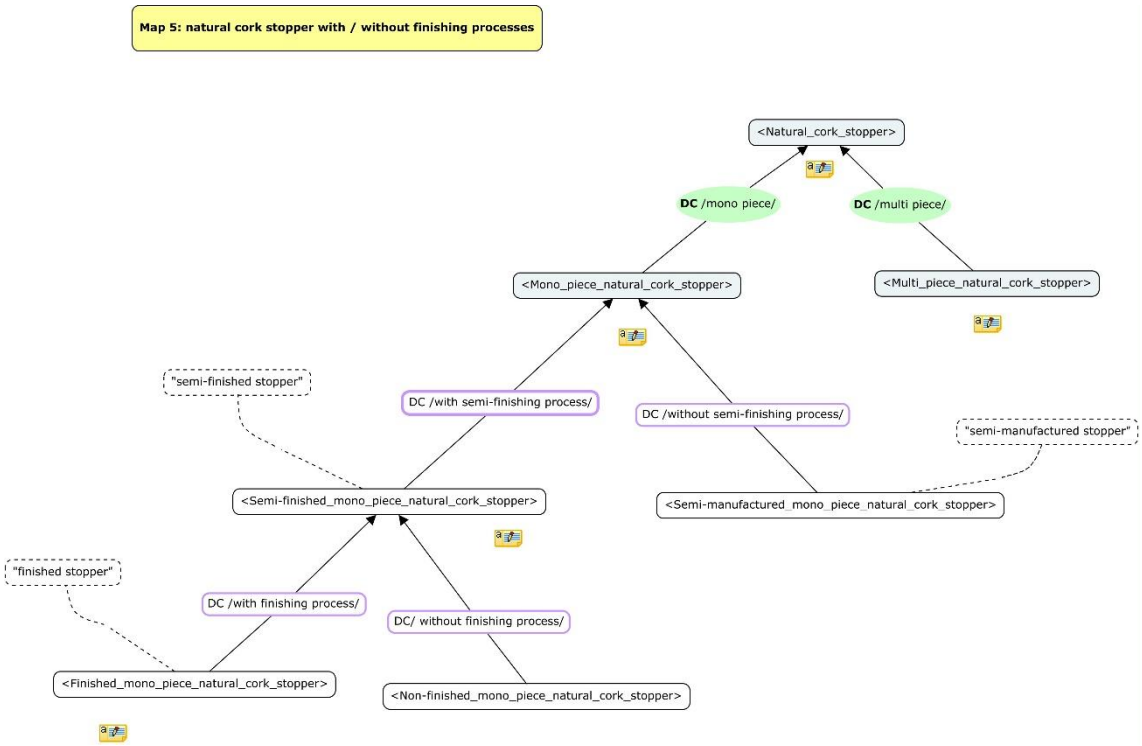
[160] *Acabamento final,* in Portuguese.

Therefore, we had to compile a collection of textual definitions drawn from the corpus for the typology of cork stoppers that were submitted to “finishing processes”; as well as a collection of textual definitions for a typology of “operations”. By typology of operations, we mean that these operations can be classified under generic concepts, namely <Semi-finishingProcesses> and <FinishingProcesses>. This typology of operations was hierarchically systematised with CmapTools (see Annex 5) and then inserted in the ontology.

The ontological modelling of these operations is further addressed in Section 6.3.5. (p. 250).

6.3.1. Finishing process or not: a differential characteristic modelled with complex axioms

Before stepping into formal modelling the hierarchy of operations that a stopper may undergo, we decided to initiate the task with a non-formal format. We have thus elaborated Conceptual Map 5: “natural cork stopper with / without finishing processes”, as follows:



Conceptual Map 5 – Systematisation of <NaturalCorkStopper> by virtue of the characteristics /with finishing process/; /without finishing process/ ; /with semi-finishing process/, and /without semi-finishing process/ in CmapTools.

Conceptual Map 5 above represents the systematisation of `NaturalCorkStopper` in CmapTools, where different stages of completion are possible to observe. It is modelled by differentia with the characteristics **/with semi-finishing process/**; **/without semi-finishing process/**; **/with finishing process/** and **/without finishing process/**, through which we can observe the three stages that a cork stopper may acquire during the manufacturing process depending on the types and subtypes of finishing processes. The terms that designate each of those concepts denoting the manufacturing stages of the cork stopper are graphically inserted in dashed-line balloons.

To logically classify a `CorkStopper` according to its stage, we have first added to our ontology the concept `FinishingProcesses`, a generic concept subsuming concepts denoting operations that fall under this category, as we will demonstrate in Section 6.3.5.1. (p. 252).

Concepts denoting operations are the means to formally describe the requirements for class membership – in the sense of OWL-DL – of cork stoppers that are submitted to finishing operations. Moreover, descriptions involving these same

concepts will enable the classification[161] of individuals[162] according to their manufacturing stage, like those represented above in Conceptual Map 5. To acquire such classification, we formally describe concepts subsumed by `CorkStopper` with the associative relation (owl:ObjectProperty) hasFinishingProcess and corresponding subtypes. These associative relations (types and subtypes) have domain and range restrictions in order to restrain class membership, e.g., the associative relation hasLenticelsColmationOperation has the domain property set as domain:Semi-finishedStopper and the range property set as a range:LenticelsColmation , a restriction established as a *class extension specification* (Lacy, 2005). This restriction enables the reasoner to infer – to classify – that any individual described with such binary relation, is a kind of `Semi-finishedStopper`. By binary relation, we mean "a relation between two things" (Horridge, 2011), in this case, between `CorkStopper` and `LenticelsColmation`. We will further address the topic domain-range in Section 6.4.1. (p. 266).

Thus, depending on either the involved operation or the inexistence of an operation, or even on the domain-range restrictions of the corresponding associative relation, it is possible to classify a kind of `CorkStopper` according to its different manufacturing stages, namely (i) `Semi-manufacturedStopper` (ii) `Semi-finishedStopper` and (iii) `FinishedStopper`, as we will demonstrate in the following lines.

6.3.2. The Boolean operators "or" and "not" to express the manufacturing stage

The next axiom is a general example to demonstrate how a given concept super ordinated by <CorkStopper> is described with or without <FinishingProcesses>.

(7) Subclass Of: CorkStopper
(hasFinishingProcesses some FinishingProcesses) or (not (FinishingProcesses))

[161] in the sense of reasoning, a feature of the reasoner – a plugin of the Protégé tool.

[162] "*also known as instances. Individuals can be referred to as being 'instances of classes'*" (Horridge, 2011).

Axiom expression (7) is a complex construct used to describe the characteristics underlying the specification of a `CorkStopper` regarding its state of completion. To express those characteristics, we used the Boolean operators ***or*** and ***not.***

In Manchester OWL Syntax, the Boolean operator ***or*** stands for logical disjunction and is a constructor also expressed as owl:unionOf in OWL Description Logic sublanguage[163] (OWL-DL). In example (7), the disjunction enables the description of those concepts that either pertain to the extension of concepts related to `FinishingProcesses` by virtue of the relation hasFinishingProcesses **or** those concepts that do not pertain to that extension. On the other hand, the Boolean operator not stands for the *complement*[164] property (owl:complementOf) and allows us to describe a concept "by identifying all objects that do not belong to a specified class expression (logical negation). The members of the complement class are individuals that are not in the class specified in the object of statement." (Lacy, 2005, p. 230).

With the Boolean operator ***or***, a differentiation is therefore set in axiom (7), so that the reasoner classifies concepts that satisfy membership in either of those two classes of concepts mentioned above, depending on their characteristics within the axis of analysis *hasProcess.*

Hence, construct (7) asserts that concepts falling under the generic `CorkStopper` may be described:

(i) through the relation hasFinishingProcesses, which corresponds to the characteristics /with finishing process/ and /with semi-finishing process/ along sub-concepts of `FinishingProcesses`;

or (ii) through the absence – the negation – of such relationship, corresponding to the characteristics /without finishing process/ and /without semi-finishing process/.

---

[163] "The primary purpose of the OWL DL sublanguage is to provide a Description Language (DL) dialect that supports reasoning applications." (Lacy, 2005, p. 138).

[164] According to Rosen, "The complement of the set A is the set $\overline{A} = U - A = \{ x \mid x \notin A \}$ containing every object not in A, where the context provides that the objects range over some specific universal domain U." (2000, p. 56).

The class of concepts described with the absence of those characteristics is the set of concepts corresponding to the *complement* – in the sense of set theory – of `FinishingProcesses`. Thus, `CorkStopper` not only subsumes concepts related to sub-concepts of `FinishingProcesses` along the relation hasFinishingProcesses, and corresponding subtypes of relations, but also subsumes concepts that do not relate through that associative relation. The latter are those concepts that denote the stage of "semi-manufacture" – a topic discussed in the next Section (6.3.3.).

To sum up, with the 3 Boolean operators used in one single axiom, namely (i) **some**, (ii) **or**, and (iii) **not**, we are thus describing that a given concept pertaining to the extension of `CorkStopper` may have **at least one** `FinishingProcesses` **or none**, by virtue of the associative relation of hasFinishingProcesses and corresponding sub-types. Therefore, we are stating that there are two kinds of `CorkStoppers` in the ontology: those with finishing treatments and those without.

6.3.3. Semi-manufacturedCorkStopper, the goal of the operator "not"

The option of including the notion of *complement* (owl: complementOf) in the description of the generic concept `CorkStopper` aims at the classification of kinds of `CorkStopper` according to their semi-manufactured stage. At this stage, a <Cork stopper> is classified as <Semi-manufactured cork stopper> if it is a kind of <Cork stopper> **without any** sort of finishing process. We have thus included in the ontology a class of concepts labelled `Semi-manufacturedCorkStopper` in order to enable the classification of concepts accordingly, i.e., whenever concepts **are not described** by characteristics along with the relation hasFinishingProcesses.

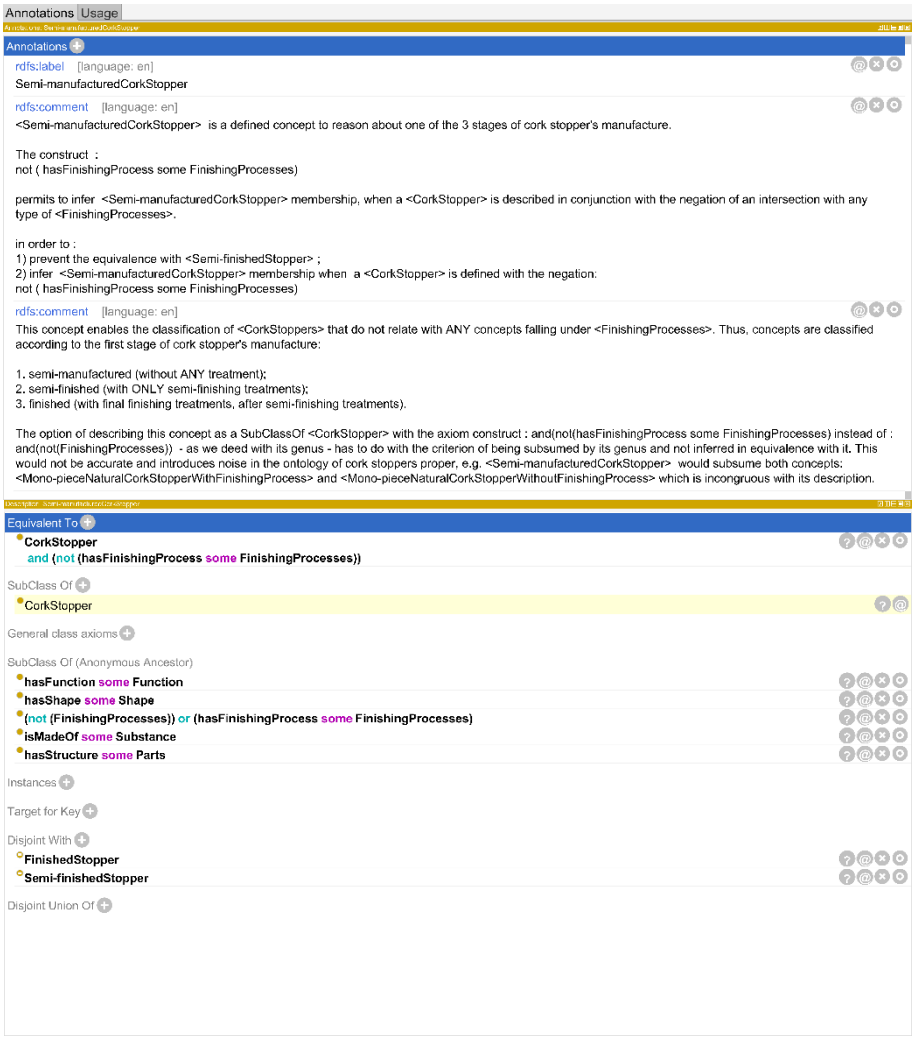


Figure 44: Description of <Semi-manufacturedCorkStopper>, a sub-type of <CorkStopper> within which the axiom construct **and (not (hasFinishingProcess some FinishingProcesses))** *is the differential characteristic from its proximum genus.*

Figure 44 shows the description of `Semi-manufacturedCorkStopper` in Protégé. It is a *defined concept*[165] and has a complex axiom, as we can observe under Equivalent To. With this axiom, we are creating a class restriction that asserts that all individuals belonging to the extension of `CorkStopper` are a kind of `Semi-manufacteredCorkStopper` if a necessary condition for membership is satisfied. That is, if there is **not** an intersection between kinds of `CorkStopper` with kinds of `FinishingProcesses`.

[165] According to Napoli, "[l]es concepts peuvent être primitifs ou définis. Les concepts primitifs sont comparables à des atomes et servent de base à la construction des concepts définis, c'est-à-dire qui possèdent une définition. À l'image d'un concept, un rôle peut être primitif ou défini et peut posséder une description structurée, où figurent les propriétés associées au rôle" (Napoli, 1997, p. 8).

This condition is replicated bellow in Example (8):

(8) SubClass Of: CorkStopper
and (not (hasFinishingProcesses some FinishingProcess))

The translation of Example (8) in natural language is: all cork stoppers that do not have any kind of finishing processes.

The notion of *do not have any kind of finishing process* is acquired through the negation constructor "not (hasFinishingProcesses some FinishingProcess)". Thus, when satisfying this condition, concepts are classified as `Semi-manufacturedCorkStopper`, as we can observe highlighted in yellow in Figure 45 below: the reasoner HermiT[166] has inferred from the intersection (and) between `MonoPieceAgglomeratedCorkStopper` and the class restriction (not (hasFinishingProcesses some FinishingProcesses)) that a `MonoPieceAgglomeratedCorkStopperWithouFinishingProcess` is a kind of `Semi-manufacturedCorkStopper`.

[166] HermiT is a plug-in of Protégé: it "is [a] reasoner for ontologies written using the Web Ontology Language (OWL). Given an OWL file, HermiT can determine whether or not the ontology is consistent, identify subsumption relationships between classes, and much more." (See http://www.hermit-reasoner.com/).

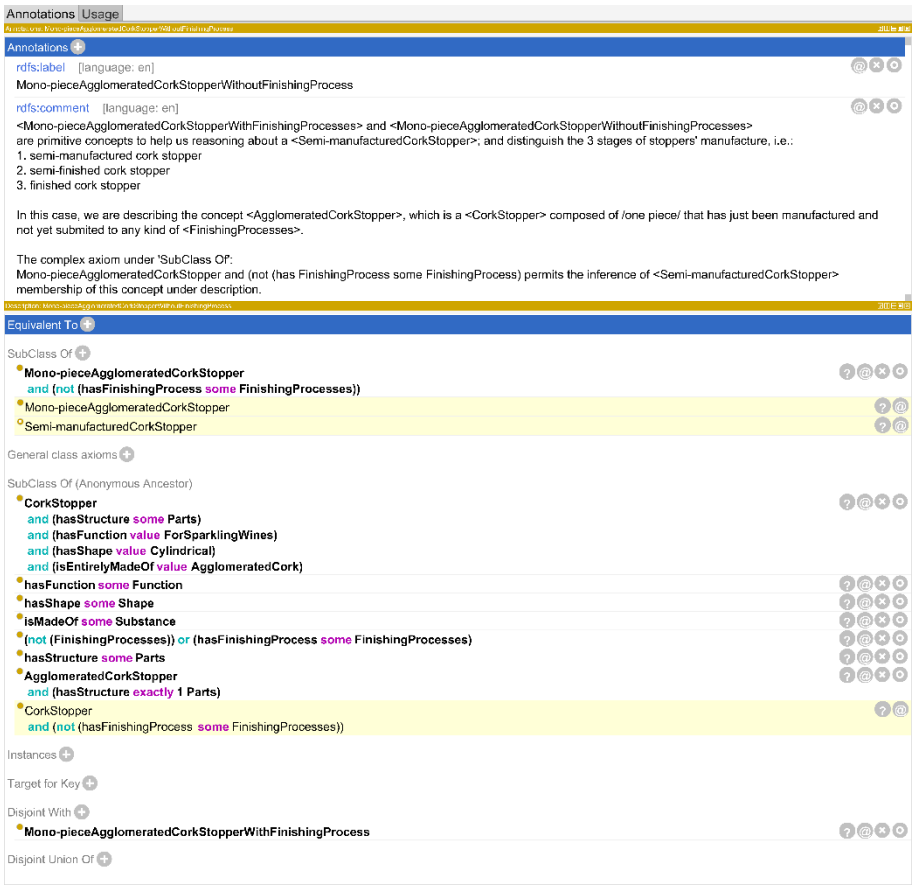


Figure 45: Classification of a kind of <CorkStopper> according to its manufacturing stage, in this case, a <Semi-manufacturedCorkStopper> given the differential characteristic conveyed by the axiom construct (not(hasFinishingProcess some FinishingProcess)).

As previously mentioned, the negation constructor (not (FinishingProcesses)) is a class restriction known as *complement* in Set Theory, but in Manchester owl Syntax – the syntax used in Protégé – it is expressed through the operator ***not***, an element that is a "logical connective used in place of ¬ " (Rosen, 2000, p. 41). We can observe its corresponding OWL constructor in the OWL-XML schema below – an excerpt of the OWL file of the ontology – for `Semi-manufacturedCorkStopper`:

```
EquivalentClasses
    Class IRI="#Semi-manufacturedCorkStopper"/
    ObjectIntersectionOf
      Class IRI="#CorkStopper"/
      ObjectComplementOf
        ObjectSomeValuesFrom
          ObjectProperty IRI="#hasFinishingProcess"/
          Class IRI="#FinishingProcesses"/
        /ObjectSomeValuesFrom
      /ObjectComplementOf
    /ObjectIntersectionOf
  /EquivalentClasses
```

6.3.4. The Boolean operators and the plurality of syntaxes to express them

As observed in the last examples, namely the OWL-XML schema to demonstrate the property ObjectComplementOf, whose corresponding format in Manchester OWL syntax is the operator ***not***, there is a plurality of ways to express the same property.

We will briefly address this topic in the following sections to support the terminology and further axiom constructs.

Let us take, for instance, the operator some. This operator, in Manchester OWL syntax, stands for "Existential Restriction" a constructor also expressed as owl:someValuesFrom, as mentioned before. However, in a more complex abstract syntax of Description Logics (DL) proper, such as $\mathcal{ALC}$[167], the existential restriction used for concept relation restrictions may be expressed as $\exists r.C$ (for r=relation, and C=concept) (see Baader, Horrocks, Lutz, & Sattler, 2017, p. 12). This plurality of syntaxes mirrors different levels of user expertise in the field of DL in close connection with their application or purpose. Thus, depending on the environment or recipient – human or machine – a different syntax may be observed.

For non-experts of DL, such plurality of syntaxes requires a sort of translation between each format for the sake of human readability. Not only because of their ability to express the same information but also due to the various formats of constructors we have found in both OWL-XML files and the Protégé tool class editor. Thus, a short example of this expressivity is shown.

Table 29: The Manchester OWL Syntax OWL 1.0 Class Constructors, based on (Horridge, et al., 2006) and corresponding Class Constructors on DL syntax and Manchester OWL syntax.

OWL Constructor	Description Logic syntax	Manchester OWL syntax	Example
intersectionOf	$C \sqcap D$	C and D	Animal and Rational
unionOf	$C \sqcup D$	C or D	Father or Mother
complementOf	$\neg C$	not C	not Quadruped
oneOf	$\{a\} \sqcup \{b\}$ …	{a b …}	{Ringo Paul John George}
someValuesFrom	$\exists$ R C	R some C	hasChild some Child
allValuesFrom	$\forall$ R C	R only C	hasChild only Female
minCardinality	$\geq n$ R	R min 3	hasChild min 3

[167] Attribute Language with General Complement. For more details see (Baader, Horrocks, Lutz, & Sattler, 2017).

maxCardinality	$\leq n$ R	R max 3	hasChild max 3
cardinality	$= n$ R	R exactly 3	hasChild exactly 3
hasValue	∃ R a	R value a	hasChild value Mary

The Manchester OWL syntax was created for OWL ontology editing tools such as Protégé and has a format easily human readable, as we can observe above, in Table 29. According to its authors, it is a formal language that was "*developed in response to a demand from a wide range of users, who do not have a Description Logic background, for a "less logician like" syntax."* (Horridge, et al., 2006).

Based on this plurality of syntaxes to describe a given concept, we can formally express the XML-DL schema (shown before)

```
EquivalentClasses
    Class IRI="#Semi-manufacturedCorkStopper"/
    ObjectIntersectionOf
      Class IRI="#CorkStopper"/
      ObjectComplementOf
        ObjectSomeValuesFrom
          ObjectProperty IRI="#hasFinishingProcess"/
          Class IRI="#FinishingProcesses"/
        /ObjectSomeValuesFrom
      /ObjectComplementOf
    /ObjectIntersectionOf
  /EquivalentClasses
```

in one single axiom using an *equality axiom* (see (Baader & Nutt, 2003), in Description Logic syntax. Thus, `Semi-manufacturedCorkStopper` can be defined as:

Semi-manufacturedCorkStopper ≡ CorkStopper ⊓ ∃hasFinishingProcess. (¬ FinishingProcess)

In DL, equality axioms are said to express definitions: *"An equality whose left-hand side is an atomic concept is a definition. Definitions are used to introduce symbolic names for complex descriptions."* (see Baader & Nutt, 2003, p. 55).

We will not address this syntax in too much detail beyond what was summarised in Table 29. We have used such syntax merely to translate into one single axiom, a previously described concept.

As a conclusion of this section, we would like to highlight the operators shown in the equality axiom above. These single axioms (or concept descriptions) are also expressed with the Boolean constructors, but graphed with mathematic symbols: conjunction (⊓), which is interpreted as set intersection; disjunction (⊔), which is interpreted as union; negation (¬), which is interpreted as a set complement; the existential quantifier restriction constructor (∃R.C); the universal quantifier restriction constructor (∀R.C), just to name a few. For more details, see Baader, Horrocks, and Sattler (2009).

The construct axioms used throughout our ontology construction chiefly focus on the most common mathematical (logical) operators for concept composition, i.e., for concepts description and subsequent membership to a given class of concepts by means of their sufficient and necessary conditions, in the sense of DL in Set theory.

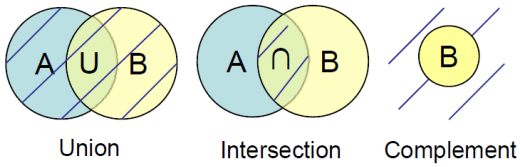


Figure 46: Illustration of Union; Intersection and Complement operators in set theory in (Lacasta, Nogueras-Iso, & Zarazaga-Soria, 2010).

For a better visualisation of how those Boolean operators work as concepts constructors, we have resorted to the relationship modelling of "composition operators" as proposed by Lacasta, Nogueras-Iso, and Zarazaga-Soria (2010) shown in Figure 46 above. According to Lacasta et al.,

> [T]he intersection composition operator […] covers practical mapping requirements. It is used to create concepts whose meaning is restricted to the common elements of two (or more than two) other concepts. For example, the concepts animal and biology can be combined to create the animal biology concept of GEMET; then this concept can be used to classify the records that are about both of the original subjects. The set of records classified according to this new concept would be the intersection of those classified with animal and those with biology. (2010, p. 31)

After this notion of concept composition, intersection – interpreted as conjunction (⊓) – is one of the most relevant operators in ontology mapping, not just due to the fact that it is implied by default in the relation of subsumption, but also to the extent that it provides the classification of individuals that are described by the intersection of two or more concepts.

In this line of thought, and following our interest in the concepts `Semi-manufacturedCorkStopper`, `Semi-finishedStopper` and `FinishedStopper`, we have pinpointed these three concepts in the ontology as concepts that denote manufacturing stages for the classification of individual membership to their extensions. To accomplish this, we have resorted to axiomatic constructs built with (i) the *complement* operator for the first and (ii) the intersection, for the other two, as demonstrated in the next Sections.

6.3.5. The extension of FinishingProcesses

We will now focus on the extension of the generic concept `FinishingProcesses`. As mentioned before, this generic concept subsumes 21 concepts. These 21 concepts play a critical role in the classification of cork stoppers according to their manufacturing stage.

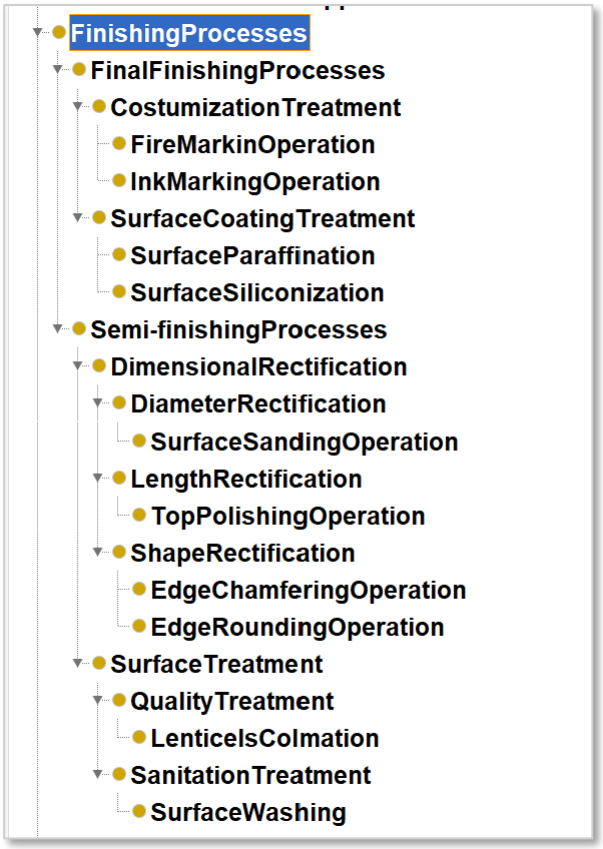


Figure 47: The extension of the concept <FinishingProcesses> in Protégé.

Figure 47 is the extension of `FinishingProcesses`. This hierarchy was not built by means of *differentia* specification all the way down – from genus to species – unless for the two first sub-types, namely `FinalFinishingProcesses` and `Semi-finishingProcesses`. These two concepts are meant to denote a sort of differentiation incremented with a touch of a temporal dimension. That is, like the information pointed at by the names of these two concepts, one occurs after the other in a timeline. As for the remaining sub-concepts, we decided instead to analyse the information conveyed by the few natural language definitions found in the corpus and systematise it according to a set of pre-established criteria; a different systematisation from what we have hitherto adopted. The details of those criteria are addressed in the next Section.

6.3.5.1. Systematisation of concepts falling under the category of FinishingProcesses

The following systematisation approach may introduce a hybrid aspect to the ontology, despite denoting an unorthodox methodology in view of the best practices for building ontologies – such as the "requirement that all differentiating notions in each part of the primitive skeleton be of the same sort." (see Rector, 2003, p. 3) – the outcomes if this ontology fulfils our expectations, when it comes to membership classification and to most of the competency queries, as discussed in the following lines.

In our ontology, `FinishingProcesses` is a generic concept we have created to subsume 21 concepts that correspond to the activities pertaining to the finishing process of cork stopper manufacturing. For the description of this subsumption, we will start from the most generic concept until we stop in the most specific one.

`FinishingProcesses` subsumes two types of operations:

(i) the final finishing operations and (ii) the semi-final finishing operations.

Each of these two types subsumes concepts that fulfil the conditions of inclusion. These conditions, in turn, are a set of criteria we had to design previously.

The criteria used for the systematisation of `Semi-finishingProcesses` are based on two considerations: the location where the operation is performed – edges; tops; entire surface; lateral surface – and the purpose of the operation – rectifying the shape or treating the surface – thus resulting on the following hierarchy:

Dimensional rectification operations

Purpose: Shape rectification

Location: Edges

term: *chamfering* – concept: EdgeChamferingOperation

term: *rounding* – concept: EdgeRoundingOperation

Purpose: Diameter rectification

Location: lateral surface

term: *surface sanding*[168] – concept: SurfaceSandingOperation

[168] "ponçagem", in Portuguese.

Purpose: Length rectification

Location: Tops

term: *top polishing*[169] – concept: TopPolishingOperation

Surface treatment operations

Purpose: Quality treatment

Location: lenticels / entire surface

term: *colmation* – concept: LenticelsColmation

Purpose: Sanitization treatment

Location: entire surface

term: *washing* – concept: SurfaceWashing

On the other hand, the criteria for the systematisation of `FinalFinishingProcesses` are based on two other considerations: (i) the operations that occur on the entire surface of the cork stopper and (ii) those occurring on different parts of the cork stopper's surface, i.e., solely on either the top or the lateral surface. Hence, the following hierarchy:

Final Surface treatment operations

Purpose: Surface coating treatment (quality improvement)

Location: entire surface

term: *paraffination* – concept: SurfaceParaffination

term: *siliconization* – concept: SurfaceSiliconization

Purpose: Customisation treatment

Location: top (s)

term: *fire marking* – concept: FireMarkingOperation

Location: Side surface

term: *ink marking* – concept: InkMarkingOperation

[169] "topojamento", in Portuguese.

As a final remark, the expert, when writing, does not make a clear difference between “fire marking” and “ink marking” in normative texts. Both concepts are randomly designated by the terms “marking” or “branding”. Nonetheless, we decided to separate them in our ontology, as an outcome from our linguistic analysis of the definitional contexts and subsequent mapping in CmapTools. The outcome of this mapping is a conceptual map that can be visualised on Annex 5. It was the starting point of the denomination process for the associative relations within the axis of analysis *hasProcess.*

Thus, as an outcome of our systematisation criteria, we can observe that the purpose of the operation occupies a more generic place in the hierarchy, while the location of the operation occupies a more specific place. There are very few definitions of each of these processes, and the few definitions we managed to draw from the CorkCorpus solely focus on one of the following three aspects: (1) the major rectification, (2) the treatment's location, or merely (3) the purpose of the treatment. This is the reason behind the systematisation of some concepts that only have one sub-concept, in the ontology, i.e., the genus-concept denotes the *purpose* of its unique species-concept that, in turn, denotes a *treatment*.

6.3.6. A competency question to validate the systematisation: what is an InkMarkingOperation?

We believe that this hierarchical systematisation of `FinishingProcesses` is relevant to understand what the purpose of each treatment is, without the need for (formally) defining each of them.

When questioning Protégé about what a given operation/treatment is, one can see, for instance, through the hierarchy shown on the tool’s answer, that an `InkMarkingOperation` is a `CustomizationTreatment` which, in turn, is a

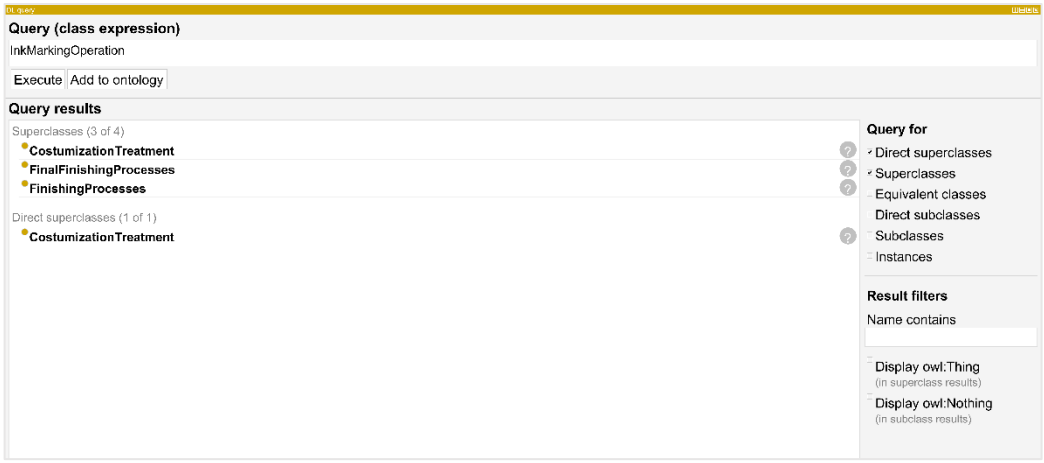


Figure 48 : Competency question in Protégé: What is an Ink marking operation?

`FinalSurfaceTreament`; and all three of them are kinds of `FinishingProcesses`, as we can observe in Figure 48 above. This is an answer that satisfies our ontological goals: it coherently states, for instance, what the purpose of such operation is, in addition to acknowledging which stage of the finishing process it belongs to.

However, we have observed, when questioning this systematisation – a subsumption within which none of the concepts is described with an axiomatic construct – one of the tools' answers lacks accuracy from the perspective of the hierarchical order, as demonstrated in the next Figure.

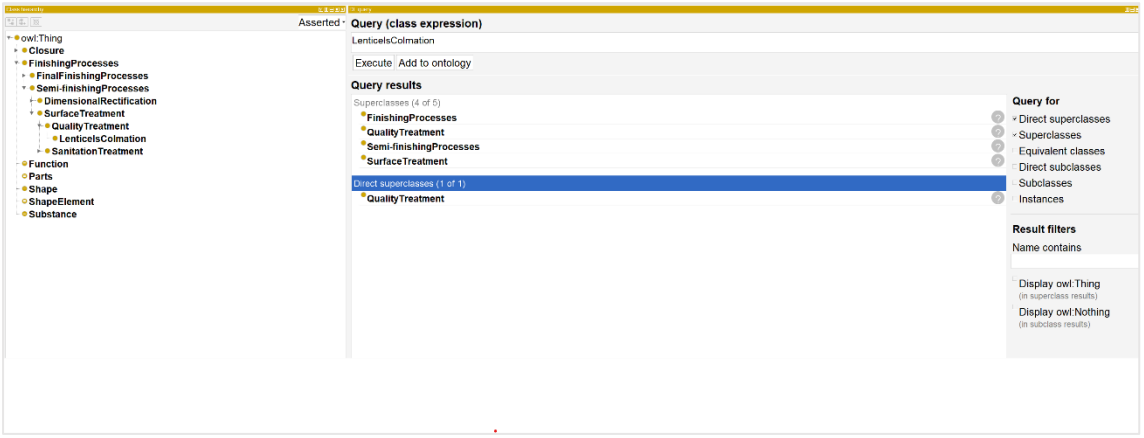


Figure 49: Competency question in Protégé: What is Lenticels colmation?

As we can see in Figure 49, if we ask what `LenticelsColmation` is, the tool displays `QualityTreatement` as being its *direct superclass*, which is perfectly accurate. However, when looking at the location occupied by `QualityTreatement` in

the hierarchy of all *superclasses*, this concept appears as a direct subclass of `FinishingProcesses` – the top of the hierarchy – instead of occupying the lowest level at the bottom of *superclasses.* In doing so, it would correctly demonstrate that it is a species of `SurfaceTreatment`. This one would demonstrate it is a species of `Semi-finishingProcesses`. And finally, the last one would demonstrate it is a species of `FinishingProcesses`. Hence, the order of concepts shown in Figure 49 above does not accurately match their order within the extension of `FinishingProcesses`. This issue consequently inserts some noise in the answer to the extent that one does not visualise the subsumption in the correct sequence.

We did not find a solution to this issue. Despite having reviewed the hierarchy and referred to best practices for ontology building, such as disjoining all siblings, the issue was not solved. This shortcoming led us to question the possibility of being a sort of limitation of the reasoner – a question that could be discussed with the tools' creators at a later stage of this project.

6.3.7. What is a CorkStopper with FinishingProcesses?

So far, we have discussed the criteria for the systematisation of the `FinishingProcesses` and briefly commented on how one can understand what a given operation is, in addition to what stage of the manufacturing process it belongs to. Notwithstanding, this systematisation has an ultimate goal beyond this last functionality. That goal is to provide the means to logically classify the state of completion of all kinds of `CorkStopper` related to kinds of `FinishingProcesses`, along with the relation hasFinishingProcess and corresponding subtypes.

In the following lines, we will focus on the axioms we have construed to define a `Semi-finishedStopper`.

6.3.7.1. Description of Semi-finishedStopper in Protégé

Figure 50 below represents the annotations and the description of `Semi-finishedStopper`.

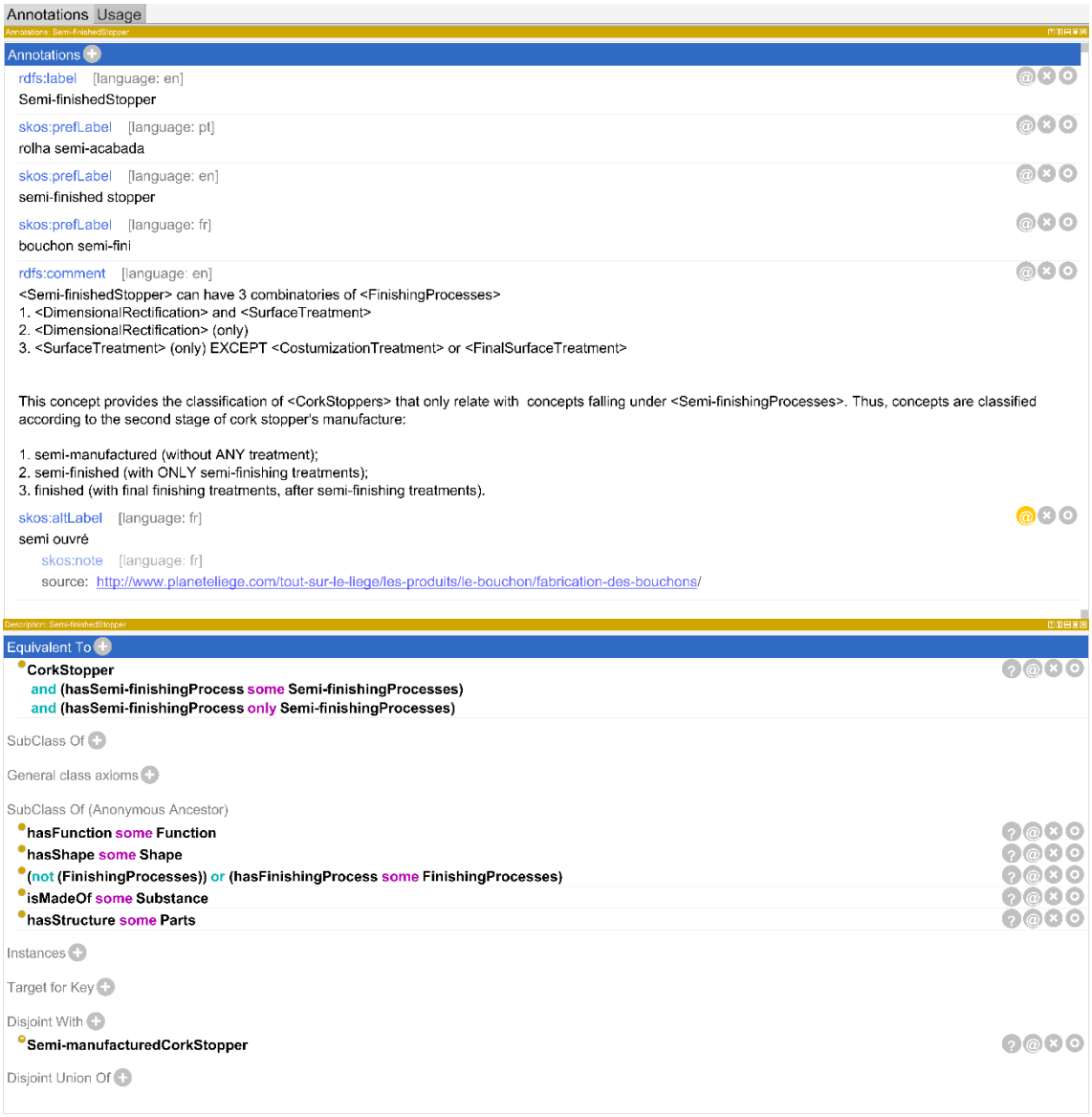


Figure 50: Description of the concept <Semi-FinishedStopper> in Protégé

As we can observe under Equivalent To in Figure 50, `Semi-finishedStopper` is a defined[170] concept described with a complex axiom. With this axiom, we are declaring that all kinds of `CorkStopper` are kinds of `Semi-finishedStopper` if the

[170] As mentioned before, when declaring necessary and sufficient conditions, we mean that the definition is complete (Section 6.1, p.219).

former satisfy the necessary and sufficient conditions to be classified as such. These conditions are put forward with a class restriction through the intersection (and) between those concepts in addition to the existential property restriction (some) along the relation hasSemi-finishingProcesses. Notice, however, the introduction of a new construct for the concept description: an additional axiom is reinforcing our statement, within which another value restriction is being used, that of ***only,*** as replicated on the following example.

(9) SubClass Of: CorkStopper
and (hasSemi-finishingProcesses some Semi-finishingProcesses)
and (hasSemi-FinishingProcesses only Semi-finishingProcesses)

The second axiom shown in Example (9) is a "*closure axiom"*, also referred to as an "*axiom restriction"* (see Horridge, 2011). Such construct requires the value restriction only, which is an operator known as the Universal restricting value property, which corresponds to the value restriction constructor (∀R.C) in DL syntax, and to the owl:allValuesFrom constraint[171] in OWL DL sublanguage. We have used this restriction to assert that no other concepts but the ones belonging to `Semi-finishProcesses` can relate to `CorkStopper` so that the latter classifies as `Semi-finishedStopper`. The rationale behind this decision has to do with the *open world assumption* (OWA)[172], an assumption that postulates that everything is true until asserted as being false. This topic will not be further discussed in our study. For more details, see Horridge (2011).

The following XML schema is an excerpt of the OWL file of the ontology for `Semi-finishedStopper`:

[171] This "constraint insists that all values for a particular property, belong to a specified class" (Lacy, 2005, p. 187).

[172] "means that we cannot assume something doesn't exist until it is explicitly stated that it does not exist. In other words, because something hasn't been stated to be true, it cannot be assumed to be false. It is assumed that 'the knowledge just hasn't been added to the knowledge base'" (Horridge, 2011, p. 63).

```
EquivalentClasses
    Class IRI="#Semi-finishedStopper"/
    ObjectIntersectionOf
      Class IRI="#CorkStopper"/
      ObjectSomeValuesFrom
        ObjectProperty IRI="#hasSemi-finishingProcess"/
        Class IRI="#Semi-finishingProcesses"/
      /ObjectSomeValuesFrom
      ObjectAllValuesFrom
        ObjectProperty IRI="#hasSemi-finishingProcess"/
        Class IRI="#Semi-finishingProcesses"/
      /ObjectAllValuesFrom
    /ObjectIntersectionOf
  /EquivalentClasses
```

Highlighted in bold, we can observe on this XML-schema both Universal and Existential value restriction constructors expressed in OWL-DL sublanguage, as well as the class restriction ObjectIntersectionOf. As seen before, these constructors correspond to the Boolean operators ***and, some***, and ***only,*** respectively, in Manchester Syntax. Here as well, we can formally express this whole information using an *equality axiom* in DL syntax:

$$\text{Semi-finishedStopper} \equiv \text{CorkStopper} \sqcap \exists \text{hasSemi-finishingFinishingProcess} \cdot \text{Semi-finishingProcess} \sqcap \forall \text{hasSemi-finishingFinishingProcess} \cdot \text{Semi-finishingProcess}$$

6.3.7.2. An example of <Semi-finishedStopper> classification

To demonstrate the classification of a given `CorkStopper` according to its semi-finished state of completion, we will take, for instance, the operation `SurfaceWashing`, a concept that belongs to the extension of `Semi-finishingProcess`.

As depicted in Figure 51 below, a `WashedMonoPieceNaturalCorkStopper` is described as a species of `MonoPieceNaturalCorkStopperWithFinishingProcess` with (the intersection) `SurfaceWashing` (along with the associative relation hasSurfaceWashing).

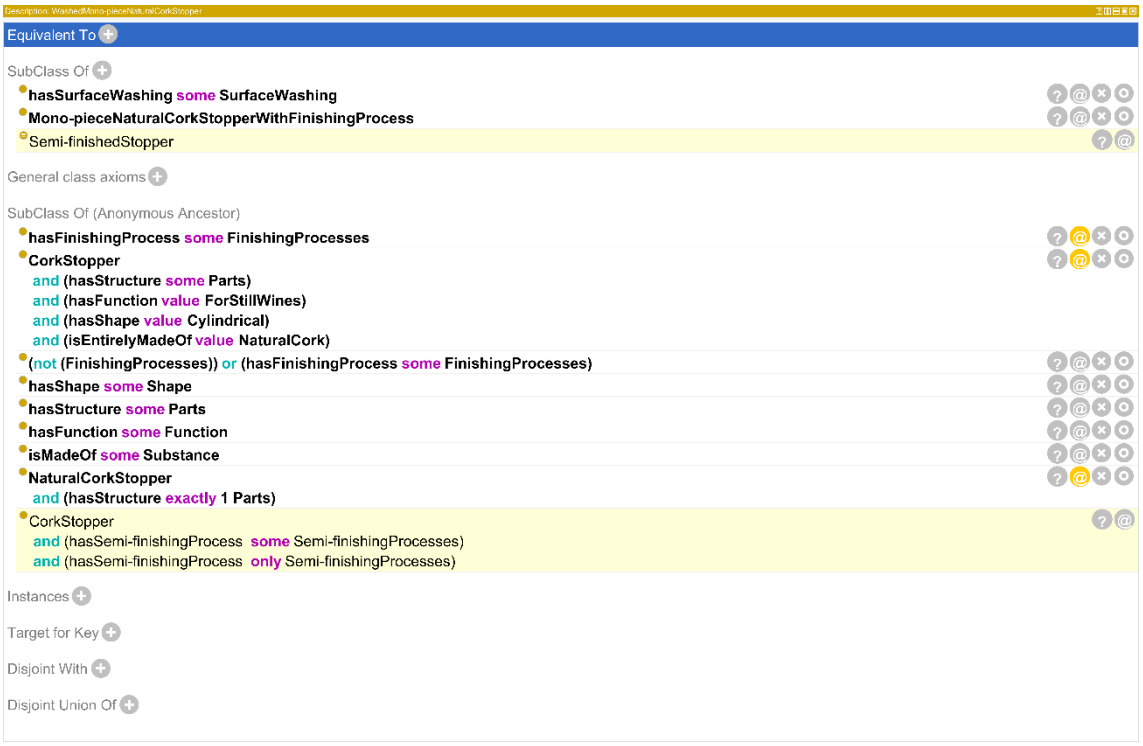


Figure 51: <WashedCorkStopper>: an example of a <Semi-finishedStopper>.

As we can see in Figure 51 above, the reasoner highlights in yellow (1) the classification and (2) the condition(s) that are satisfied for that classification:

(1) it is the equivalent concept that matches what it is being asserted by the axiom constructs, which in this case is `Semi-finishedStopper` – shown under SubClass Of, in Figure 51 – and

(2) it is the source of the inference, i.e., the axiom constructs for the description of `Semi-finishedStopper` that substantiates the inference – shown at the bottom of SubClass Of (Anonymous Ancestor) in Figure 51.

This means that the concept described satisfies one of the conditions to be classified as `semi-finishedStopper`. The condition is: hasSurfaceWashing some SurfaceWashing. Since the latter belongs to the extension of `Semi-finishingProcess`, the classification is explicitly an outcome from the restrictions we created for the description of `Semi-finishedStopper`.

6.3.8. Description of the concept FinishedStopper

In this section, we finally address the concept `FinishedStopper`. The methodology for its description is identical to the `Semi-finishedStopper`, except for the different type of relations used along with different concepts. The latter have by now been easily guessed by the reader: they are the extension of `FinalFinishingProcesses` along with their corresponding associative relations.

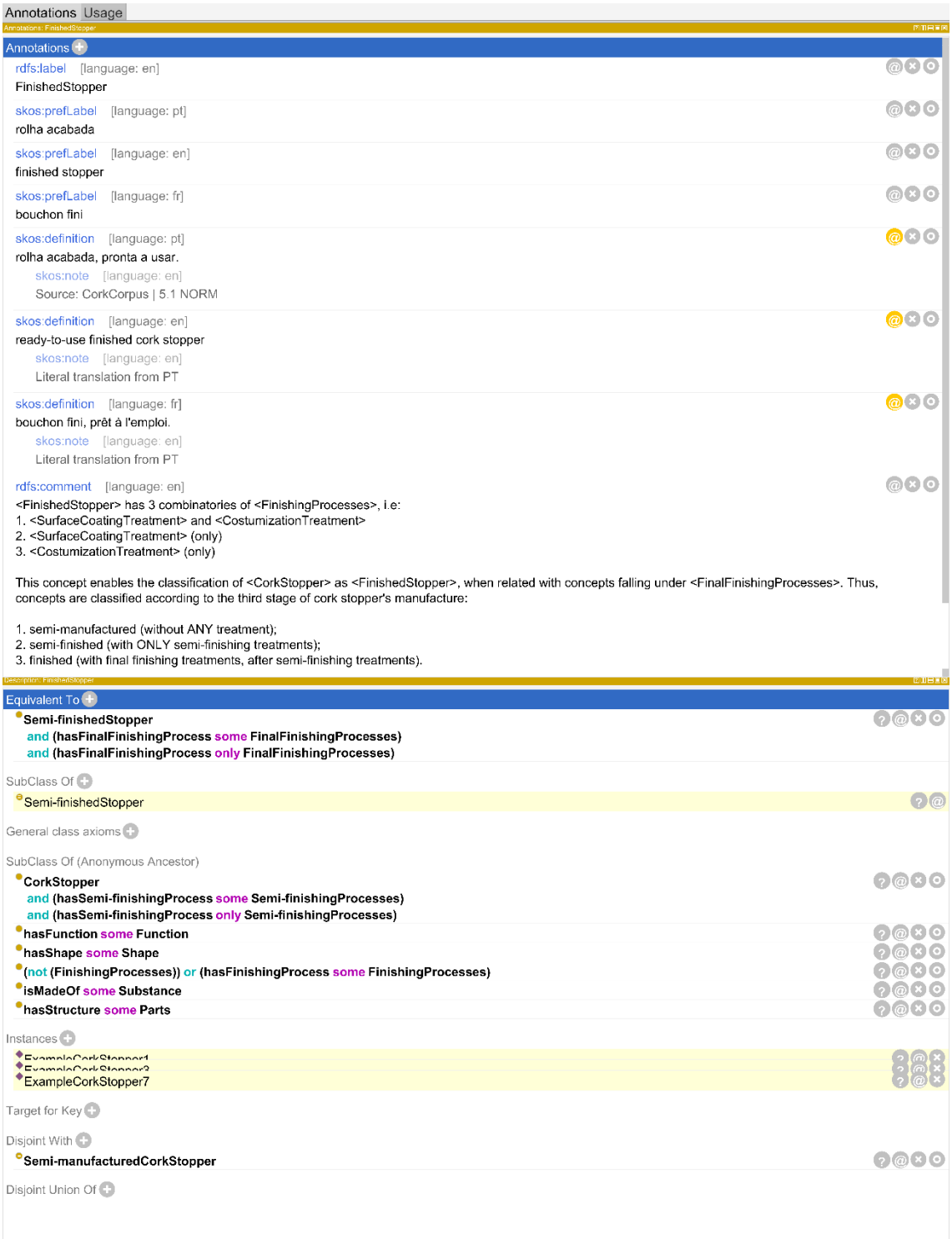


Figure 52: Description of FinishedStopper

Figure 52 corresponds to the description of `FinishedStopper`. It is a defined concept in the sense of DL, and described with a complex axiom, as we can observe under Equivalent To, highlighted in blue.

With this axiom, we are declaring that a `CorkStopper` is a `FinishedCorkStopper` if the former satisfies the necessary conditions we have stated, namely:

hasFinalFinishingProcesses some FinalFinishingProcesses
hasFinalFinishingProcesses only FinalFinishingProcesses

Once again, a closure axiom was construed to define this concept. The translation in natural language is: it must have at least one kind of final finishing process, but only final finishing process.

Finally, in DL syntax, the *equality axiom* would be:

$$\text{FinishedStopper} \equiv \text{CorkStopper} \sqcap \exists \text{hasFinalFinishingProcess} \cdot \text{FinalFinishingProcess} \sqcap \forall \text{hasFinalFinishingProcess} \cdot \text{FinalFinishingProcess}$$

Before concluding this topic, we will explore one last example, where we can see the reasoner classifying a cork stopper that has undergone a final finishing process. For that, we have chosen the `InkMarkedChamferedWashedMonoPieceNaturalCorkStopper`: a concept that denotes a "mono-piece natural cork stopper" that has been submitted to several operations, namely to:

`SurfaceWashing` → `ChamferingOperation` → `InkMarkingOperation`

in that exact order.

6.3.8.1. An example of FinishedStopper

The `InkMarkingOperation` is a customisation treatment performed on the lateral surface of the cork stopper. This operation falls under the `FinalFinishingProcess` activities, a direct sub-concept of the generic concept `FinishingProcesses`. When this operation is explicitly asserted along the associative relation hasInkMarkingOperation in a given `CorkStopper` description, such construct provides reasoning regarding concepts that satisfy the necessary and

sufficient conditions stated on the definition of `FinishedStopper` as represented below:

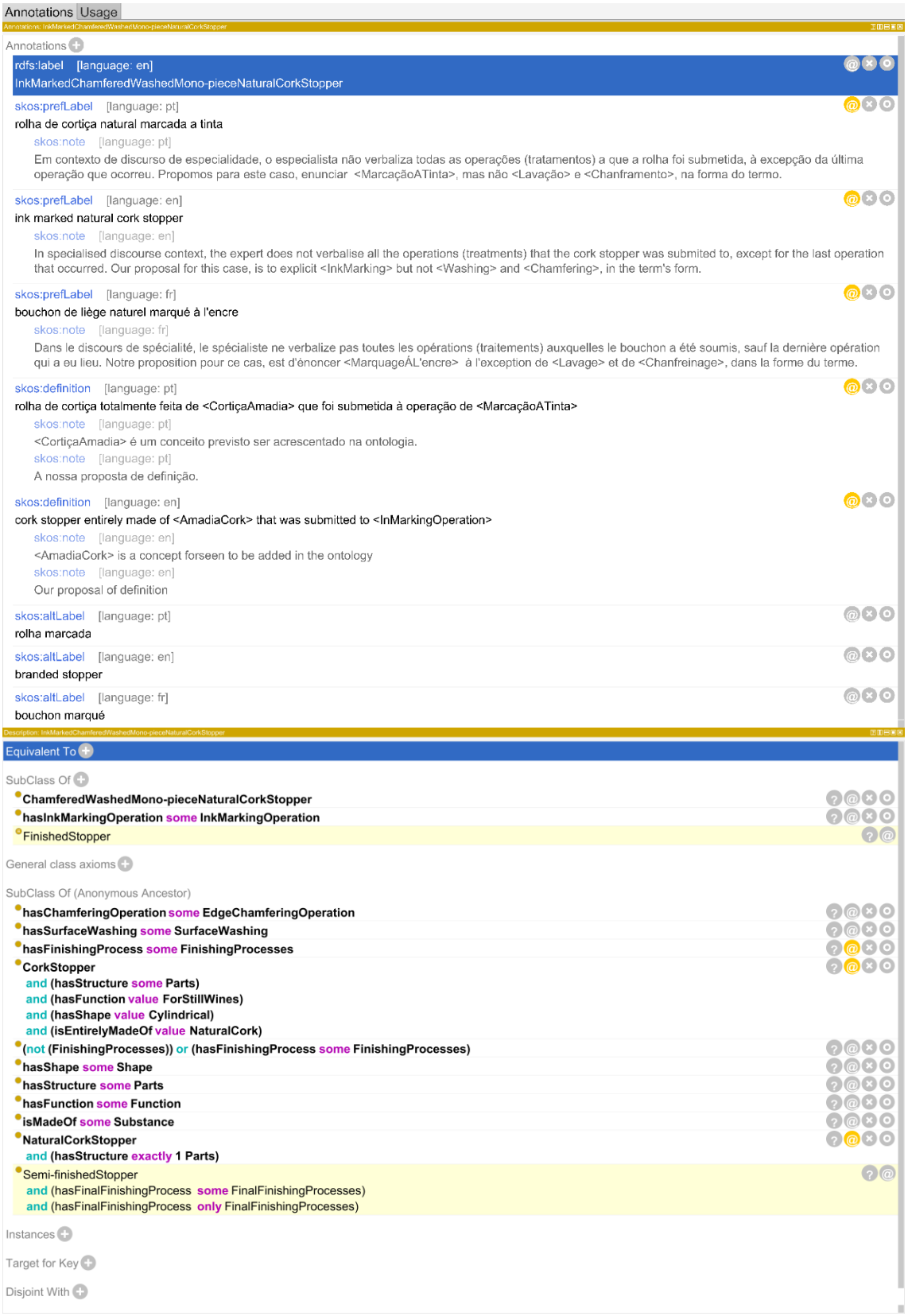


Figure 53: An "ink marked natural cork stopper" classified as <FinishedStopper>.

In Figure 53 above, highlighted in yellow, we can see the reasoner classifying an `InkMarkedChamferedWashedMonoPieceNaturalCorkStopperWithFinishing Process` – whose linguistic label is "ink marked natural cork stopper" – as a `FinishedStopper`.

This reasoning for this classification is grounded on the intersection between the relation (owl:ObjectProperty) hasInkMarkOperation and the concept `ChamferedWashedMonoPieceNaturalCorkStopperWithFinishingProcess`. As mentioned before, the intersection is implicit through the systematisation of several axioms under the SubClass Of, which in this case is:

SubClass Of:
ChamferedWashedMonoPieceNaturalCorkStopper
hasInkMarkOperation some InkMarkingOperation ⇕ "and" | ⊓

As demonstrated above, the concept here described satisfies one of the conditions to classify as `FinishedStopper`. The condition is: hasInkMarkOperation some InkMarkingOperation. Since the latter belongs to the extension of `FinishingProcess`, the classification is an outcome of the restrictions we created for the description of `FinishedStopper`.

Finally, and similarly to the transcriptions we have done of all the previous examples, this axiomatic expression in Manchester syntax corresponds to the following axiom equality, in DL syntax.

InkMarkedChamferedWashedMonoPieceNaturalCorkStopper ≡
ChamferedWashedMonoPieceNaturalCorkStopperWithFinishingProcess ⊓
∃hasInkMarkOperation · InkMarkingOperation

6.4. Hierarchical systematisation of the associative relations to relate CorkStopper and FinishingProcesses

This last section is dedicated to the associative relations we have created to relate `CorkStopper` with `FinishingProcesses`, so that we can obtain a classification of the former regarding its stage of completion in the manufacturing process.

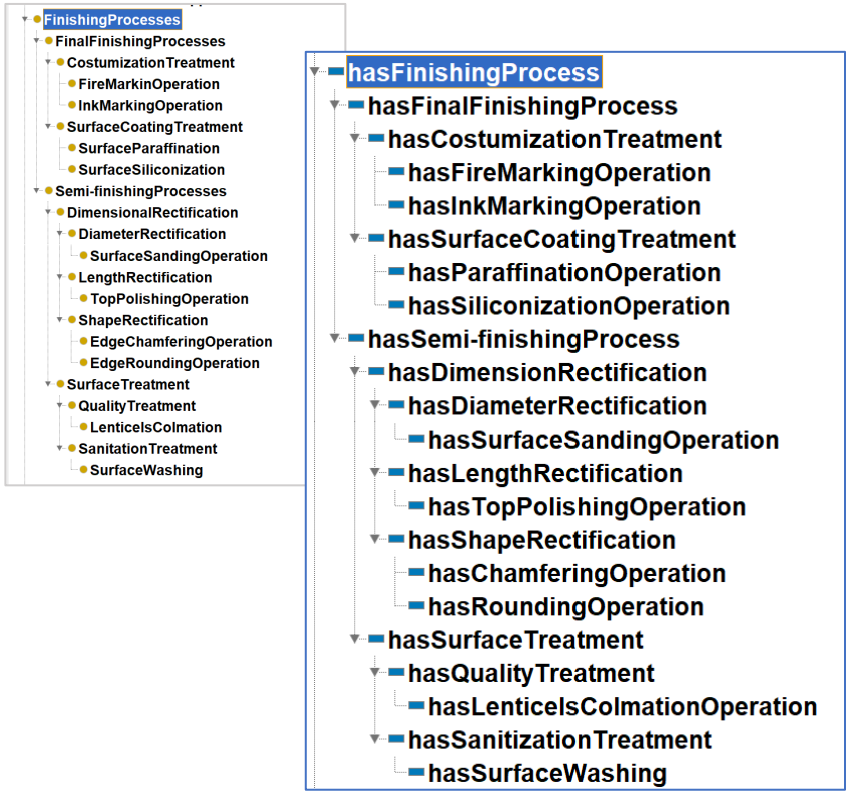


Figure 54: Owl:ObjectProperties corresponding to associative relations, sub-type [PROCESS-RESULT].

In Figure 54 above there are 2 panels: the larger panel corresponds to the hierarchical systematisation of the associative relations that relate concepts from the extension of `FinishingProcesses` and those from the extension of `CorkStoppers`, while the small panel represents the extension of `FinishingProcesses`, which was already shown in Section 6.3.5 (p. 251). As we can observe, relations (owl:ObjectProperties) are systematised in the same order as the concepts of the extension of `FinishingProcesses`.

Each of those conceptual relations is associative, subtype [PROCESS-RESULT]. The *domain*[173] of each relation (owl:ObjectProperty) is either the concept of `FinishedStopper` or `Semi-finishedStopper`, depending on the type of operation the concept denotes. However, these are not the only associative relations we needed to include in the ontology.

As mentioned before, there are concepts denoting operations that occur on a particular location of the `CorkStopper` such as the `Edge`. This last concept is one of the enumerated concepts composing the extension of `ShapeElement`; thus, populating the ontology as an *instance*[174]. The same applies to `Chamfered`, in this case, it is an instance from the extension of `Shape`. We will use these concepts, in the next Section, to address the topic of domain and range properties – a key feature to test the ontology's consistency.

6.4.1. Domain and range of the relation hasShapeElementEdge

Reasoning (i.e., running the reasoner) is an important test to ensure the quality of an ontology, not only to test its coherency – by checking if the resulting subsumption is coherent with the conditions we have stated to define the concepts – but also its logical consistency, for "Based on the description (conditions) of a class the reasoner can check whether or not it is possible for the class to have any instances. A class is deemed to be inconsistent if it cannot possibly have any instances." (Horridge, 2011, p. 48).

The need for populating our ontology with instances has a twofold purpose: firstly, it creates classes of concepts by enumeration; and secondly, it sets *domain* and *range* restrictions. These restrictions are set to the binary relation occurring between two given instances in order to provide logical constraints for their classification, as we shall demonstrate in the following lines.

---

[173] is a property that limits the use of the relation to a specified intersection of classes (see Lacy, 2005).

[174] also known as individuals. Likewise, individuals can be referred as *instances of classes* (see Horridge, 2011, p. 11). Moreover, "les entités de base qui sont définies et manipulées dans une logique de descriptions sont les concepts et les rôles. Un concept dénote un ensemble d'individus - l'extension du concept - et un rôle dénote une relation binaire entre individus" (Napoli, 1997, p. 8).

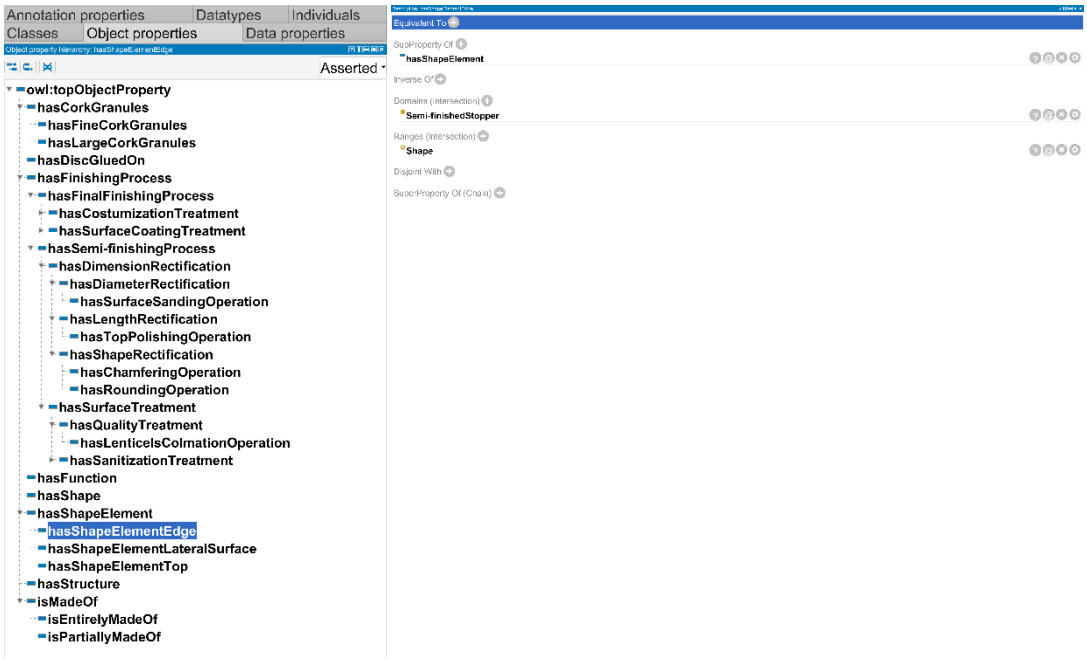


Figure 55: Domain-Range of the relation (owl:ObjectProperty) hasShapeElementEdge side by side with the whole hierarchy of conceptual relations.

Figure 55 above depicts the description of the relation (owl:ObjectProperty) hasShapeElementEdge. On the right panel, we can see the restrictions we have set to domain and range. These restrictions are the rules that underlie the classification of individuals, in this case, the ones that hold the relation hasShapeElementEdge, as described below:

(1) `Semi-FinishedStopper` (a defined concept) is the domain of the conceptual relation (owl:ObjectProperty) hasShapeElementEdge ;

(2) `Shape` (an enumerated class of concepts) is the range of the conceptual relation (owl:ObjectProperty) hasShapeElementEdge.

With this domain-range constraint, we provided rules for the classification of instances as follows:

(1.1) The subject of the statement described with the relation hasShapeElementEdge is always a kind of `Semi-FinishedStopper`. This statement leads to the classification of all instances, accordingly, if the latter are described with such relation. The classification, in turn, is obtained from the *domain* property: a property that restrains

the use of the relation to a specified intersection of subject classes (see Lacy, 2005, p. 123).

(1.2) The object of the statement described with the relation hasShapeElementEdge is always a kind of `Shape`. The relation is thus associated with members of this class of concepts. This feature is obtained from the property of *range*: a property that specifies which instance can be an object of the relation. In this case, no other instance but the ones enumerated as Shape = {Rounded; Chamfered…}.

As stated above in Item (1.1), there is a "subject of the statement", as well as an "object of the statement", this time, mentioned in Item (1.2.). These notions require further explanations since the associated topic is tied with ontological triples; a matter that we superficially address in the next section.

6.4.2. Ontological triples: a kind of declarative assertions

According to Lim, Liu, and Lee (2011), concepts holding relationships are defined as ontological triples. These triples are also called ontological statements in formal languages (e.g., RDF[175] and OWL). The definition of ontological triples is a node-graph structure, composed of two nodes (Subject and Object) and "a triple connecting them ([the]Predicate)" (W3C, 2014), as represented below, in Figure 56:

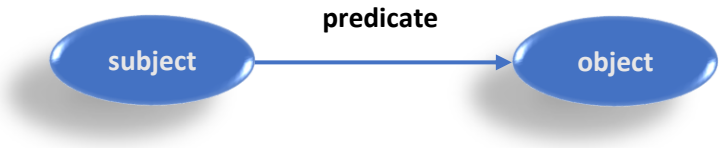


Figure 56: RDF graphs are sets of subject-predicate-object triple (in RDF 1.1 Concepts and Abstract Syntax, W3C Recommendation 25 February 2014).

A set of such triples is called an RDF graph – the core structure of the abstract syntax to link all RDF-based languages and specifications to represent information in the Web – and are used for the description of *resources*[176] (see W3C, 2014). It is interesting

[175] The Resource Description Framework (RDF) is a framework for representing information in the Web. Information available online (W3C, 2014) accessed on [22-04-2020].

[176] "any IRI (Internationalized Resource Identifier) or *literal* (lexical form; datatype IRI; or language tag) denotes something in the world (the "universe of discourse")." (*ibid.*).

to observe that such representation mirrors one of the basic orders of sentence constituents in most natural language systems – in the Saussurean sense – namely, Subject → Verb (predicate) → Object (SVO), regarding the syntactical construction of *declarative sentences* (i.e., affirmative or negative assertions). Regardless of their different fields of knowledge, namely Linguistic and Semantic Web, those two types of declarative constructs share the same functionality. Broadly speaking, when any given information is asserted through both models, a proposition is in place, in the sense of a "*declarative sentence that has a well-defined truth value"* (Rosen, 2000). Notwithstanding, in the stricter sense of the Predicate Logic perspective, the *predicate* is a declarative statement with a symbolic form: $\mathcal{P}(x)$, in which the variable (x) is not specified (see Rosen, 2000). The topic of Predicate Logic will not be addressed with more detail apart from this light explanation.

Still, either being a sentence in natural language or a statement in the broad sense of logic[177], the common feature underlying the two is *predication*, in the Aristotelian sense. As postulated by this author, when, regarding a man, you say that he is a man or that he is an animal; you are saying what the entity is and you are also implying a substance. Each of these predications, both when you say something about a thing or when you refer to its genus, you mean what the thing is (Minio-Paluello, 2016).

As a concluding remark, the emphasis here is that the two declarative statement models mentioned above share the same predicative feature of asserting *(predicating)* that something is *(signifies)* something, regardless of belonging to a formal system or to the natural language system.

6.4.3. Classification of two instances as Semi-finishedStopper

Following the above rationale, we will now represent a triple with concepts and instances from our ontology. For that, we will resort to the instance `Chamfered` – an

[177] "is the basis for distinguishing what may be correctly inferred from a given collection of facts. Propositional logic, where there are no quantifiers (so quantifiers range over nothing) is called zero-order logic. Predicate Logic, where quantifiers range of members of a universe, is called first-order logic. Higher order logic includes second-order logic (where quantifiers can range over relations over the universe), third-order logic (where quantifiers can range relations over relations), and so on. Logic has many applications in computer science, including circuit design […] and verification of computer programme correctness […]." (Rosen, 2000, p. 45).

instance that pertains to the enumerated concept `Shape` – to be the object of an RDF triple; the instance ExampleCorkStopper2, for the subject; and finally, the relation hasShapeElementEdge to hold between the two first instances, as the *predicate* of the triple.

The underlying argument here is to demonstrate the process of logical classification employing the domain and range properties.

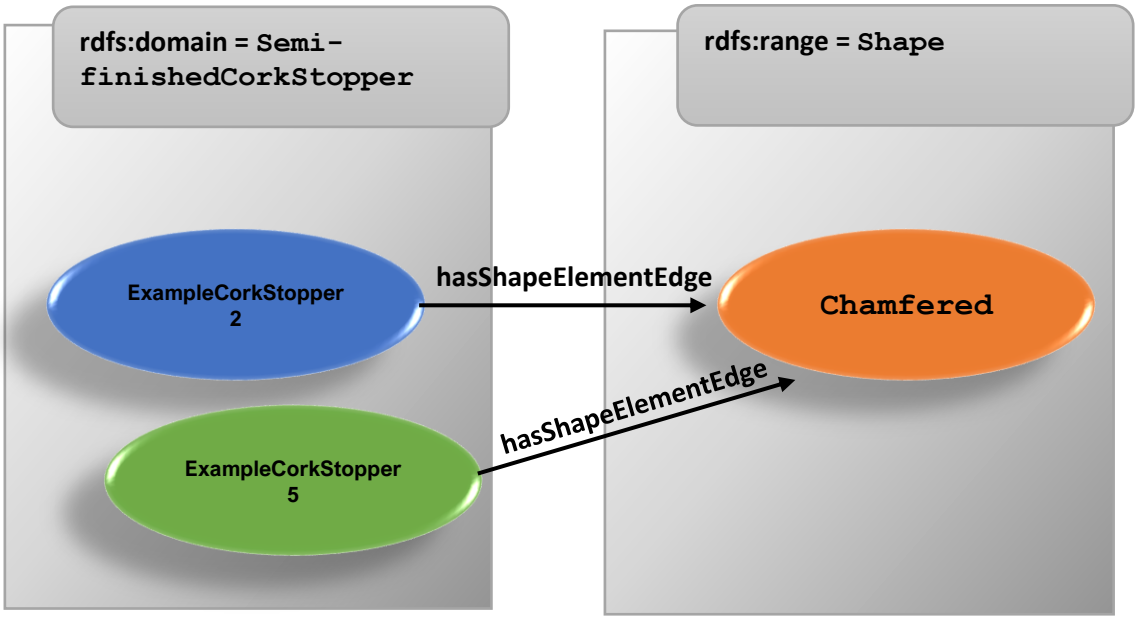


Figure 57: Classification of two instances by means of the Domain and Range properties applied to the relation hasShapeElementEdge represented in RDF triples

Figure 57 above represents the classification of two instances through the Domain and Range properties applied to the relationship hasShapeElementEdge, represented in RDF triples.

Since the domain property restrains the relation hasShapeElementEdge to be held exclusively with members of `Semi-FinishedStopper`, the instance ExampleCorkStopper2 is classified as a `Semi-FinishedStopper` because its description points at the characteristic `Chamfered` by means of the restrained relation. The same applies to ExampleCorkStopper5 given the presence of this particular characteristic in its description.

The above classification can be observed in Protégé, as follows:

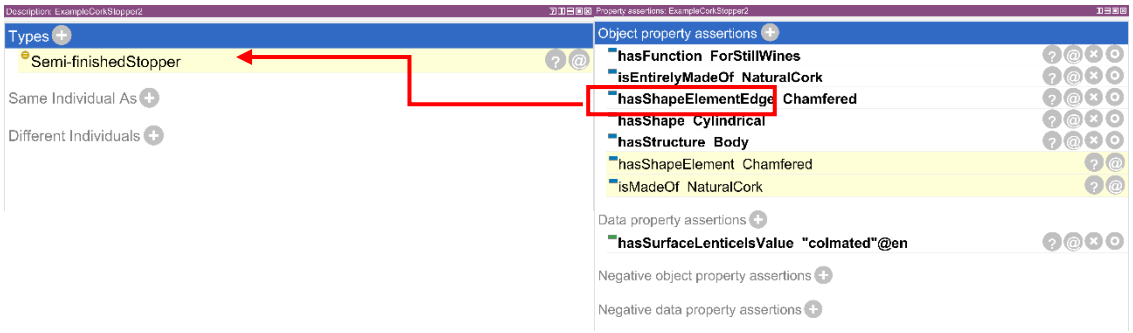


Figure 58: Classification of the instance ExampleCorkStopper2 as a <Semi-finishedStopper> by means of the domain property applied to the relation hasShapeElementEdge.

As we can see highlighted in yellow in Figure 58 above, the instance ExampleCorkStopper2 is classified as `Semi-finishedStopper` given the characteristic conveyed by the assertion hasShapeElementEdge `Chamfered`.

The same applies to all instances holding that relation, unless there is a characteristic pointed at by an associative relation that is restrained to `FinishedStopper`.

At this point, we can finally discuss the stage of `FinishedStopper` and demonstrate how the intermediate stage of `Semi-finishedStopper` shifts into the final stage of completion.

This point was one of the most challenging tasks we found while building the ontology. Given the small number of instances with which we decided to populate the ontology, namely a few concepts to compose the enumerated classes, such as `Shape`, there were not enough instances to create triples with domain and range restrictions. Thus, to continue the methodology we followed in the previous example, namely by restraining the domain of the relation hasShapeElementEdge to members of `Semi-finishingStopper`, we had to add one more enumerated class of concepts to the ontology: the `BrandMark` = {InkMark, FireMark}.

With both concepts `InkMark` and `FireMark`, and the creation of the relation hasBrandMark, whose domain is set to `FinishedStopper`, it is finally possible to get a classification of a given instance accordingly, if one of those enumerated concepts for `BrandMark` is explicit along with the relationship we have just created, as demonstrated bellow.

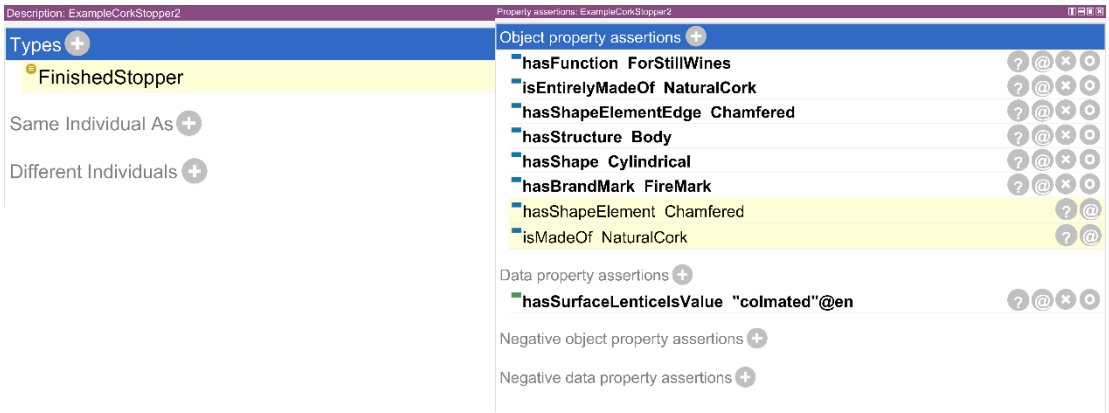


Figure 59: ExampleCorkStopper2 has an updated classification with the addition of the characteristic hasBrandMark.

As we can see in Figure 59, the description of the instance `ExampleCorkStopper2` – an instance previously classified as `Semi-finishedStopper` – has an additional assertion, namely hasBrandMark `FireMark`. This assertion corresponds to an additional characteristic. The main point here is that the classification was updated to the state of `FinishedStopper`. This shift of status is an outcome of our decision regarding `Semi-finishedStopper` subsuming `FinishedStopper`, which implies that the latter is a species of the former. This decision intended to mirror the timeline of the manufacturing process of cork stoppers: before acquiring the status of `FinishedStopper`, every stopper is at the stage of `Semi-finishedStopper`. Hence, when a given instance is described with characteristics that are conveyed by relations restrained to <FinishedStopper>, the instance gets classified as a member of that class of concepts.

As demonstrated above, the description of an instance can be as explicit as possible, as long as concepts denoting characteristics are created, along with their conceptual relations for reasoning purposes.

Nevertheless, it should be noted that a *description* does not replace a *definition*; instead, the former complements the latter. An individual's description is a description proper, in the sense of being explicative, but it is not a concept's definition. According to Rey (1979), a description may combine both pertinent characteristics and non-pertinent characteristics, but mostly non-pertinent ones, whereas a logical (formal)

definition is “constructive” and “essentielle”. The non-pertinent characteristic of Rey is what we call descriptive characteristic.

In our ontology, despite being considered as non-essential to define a concept, descriptive characteristics are useful features to assists us not only in describing an instance and obtaining a shift of its finished status, as demonstrated above, but also in classifying individuals with the same compositional structure – i.e., the same parts – yet differing given the scalable amounts of those parts, or even with the same amount of parts but located in different places. That is the case of the concept designated by the term “technical cork stoppers N+N” – our final topic, discussed in the next section.

6.5. Additional information to the definition: the case of the “technical cork stopper N+N”

The analysis of the textual definition of “technical cork stopper N+N” is not demonstrated step by step in this study. In this section, we will simply summarise our observations, which have followed the same methodology we have presented so far.

Still, some information about this peculiar cork stopper must be provided in a few lines:

The “technical cork stopper N+N” is a very special type of stopper given its composition. The “body” of the stopper is made of agglomerated cork, and each top of the “body” may have one or two “discs” glued. The “discs”, in turn, are made of natural cork. Given this double type of raw material, it is a “mixed cork stopper”, as already included in Conceptual Map 1 “Type of cork stopper”, Section 5.1.1. (p. 183). The difficulty of describing this type of stopper is tied with the triple combination of discs glued:

(i) 1+1, meaning 1 disc glued on each top;

(ii) 2+2, meaning 2 discs glued on each top; or

(iii) 0+2, meaning zero discs on one top and 2 discs glued on the other.

This explains the element “N+N” in the term’s structure, where n = digit.

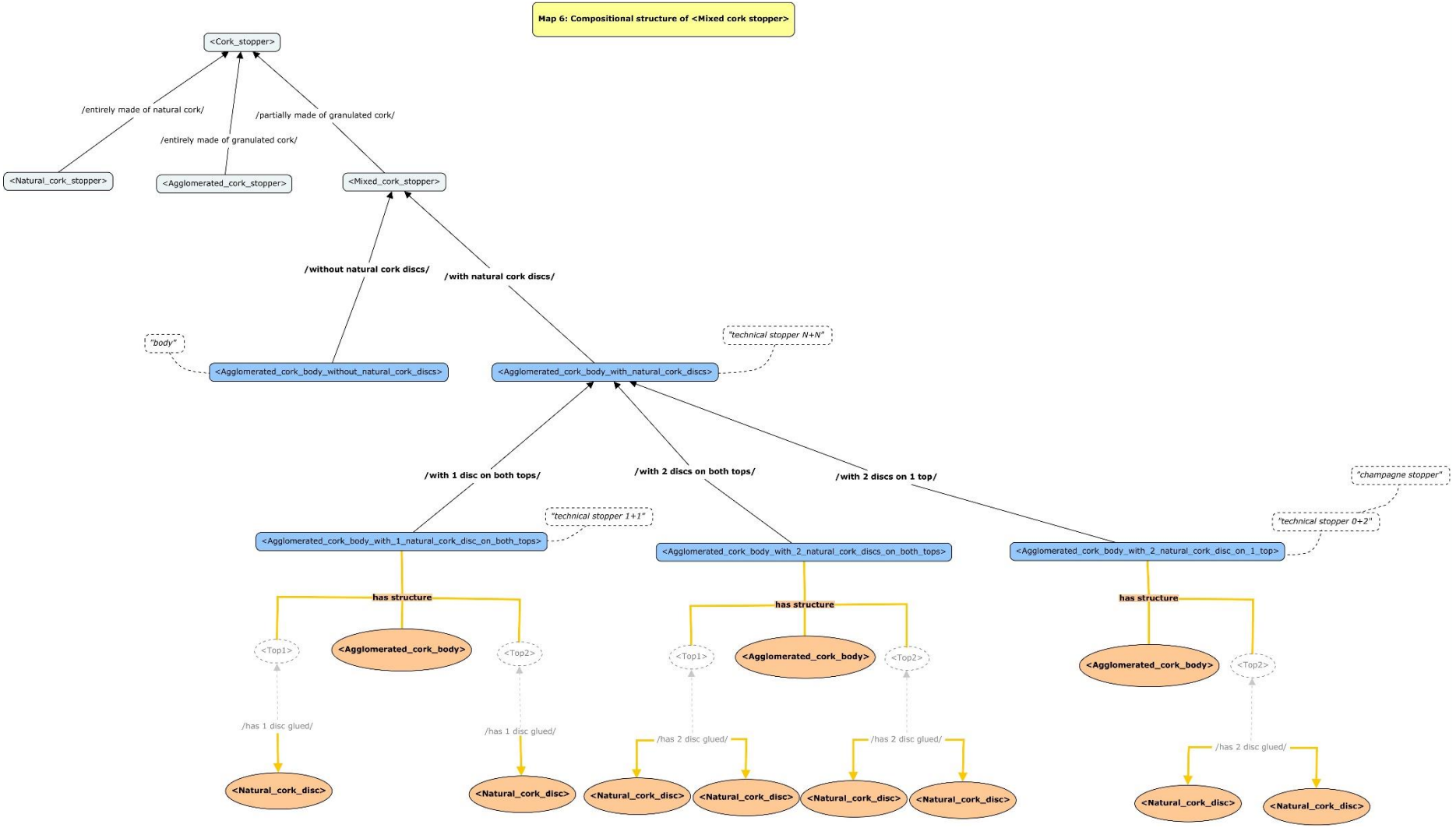
Once again, we used CmapTools to preview the knowledge organisation regarding types of "technical cork stoppers N+N″. In this particular case, CmapTools helped us to confirm that some characteristics, in particular the descriptive ones, must be made explicit through *"mixed concept systems"* (ISO 704, 2009), i.e., maps that can be simultaneously organised through different types of relations, namely generic-specific and partitive.

According to ISO,

> to identify partitive concepts and their characteristics, it is necessary to determine first the position of the comprehensive concept in a generic hierarchy and to be mindful of the inheritance principle. How generic the comprehensive concept is will determine its partitive concepts and the extension of those concepts. (ISO 704, 2009, p. 14)

As mentioned before, descriptive characteristics have to be referred, but mostly to improve the systematisation of conceptual maps and subsequently create the corresponding descriptive information in the ontology. That was the case of making explicit the location where the `Disc` – a `Part` made of `NaturalCork` – is glued on the `Body` - a `Part` made of `AgglomeratedCork`. The location is either on `Top1` or `Top2` or even both. And finally, also making explicit the number of discs glued on each of those different tops.

"Technical cork stopper N+N" is the term in natural language used to designate the concept we labelled as <Agglomerated_cork_body_with_natural_cork_discs>, as illustrated below in Conceptual Map 6.



Conceptual Map 6 – A systematisation of the compositional structure of <Mixed Cork Stopper>

As depicted in Conceptual Map 6, the partitive schematisation is coloured in orange. Although this partitive schematisation might seem redundant, this form of representation helped us to create the descriptive characteristics in the ontology, in particular regarding the concepts <Agglomerated_cork_body_with_1_natural_cork_disc_on_both_tops> and <Agglomerated_cork_body_with_2_natural_cork_disc_on_both_tops> – two of the most specific concepts coloured in blue.

Based on Conceptual Map 6, we have created several conceptual relations, in the ontology, to make explicit both essential and descriptive characteristics: the former for the concept description and the latter for the description of an instance, as we shall demonstrate in the last section.

6.5.1. Descriptive characteristics

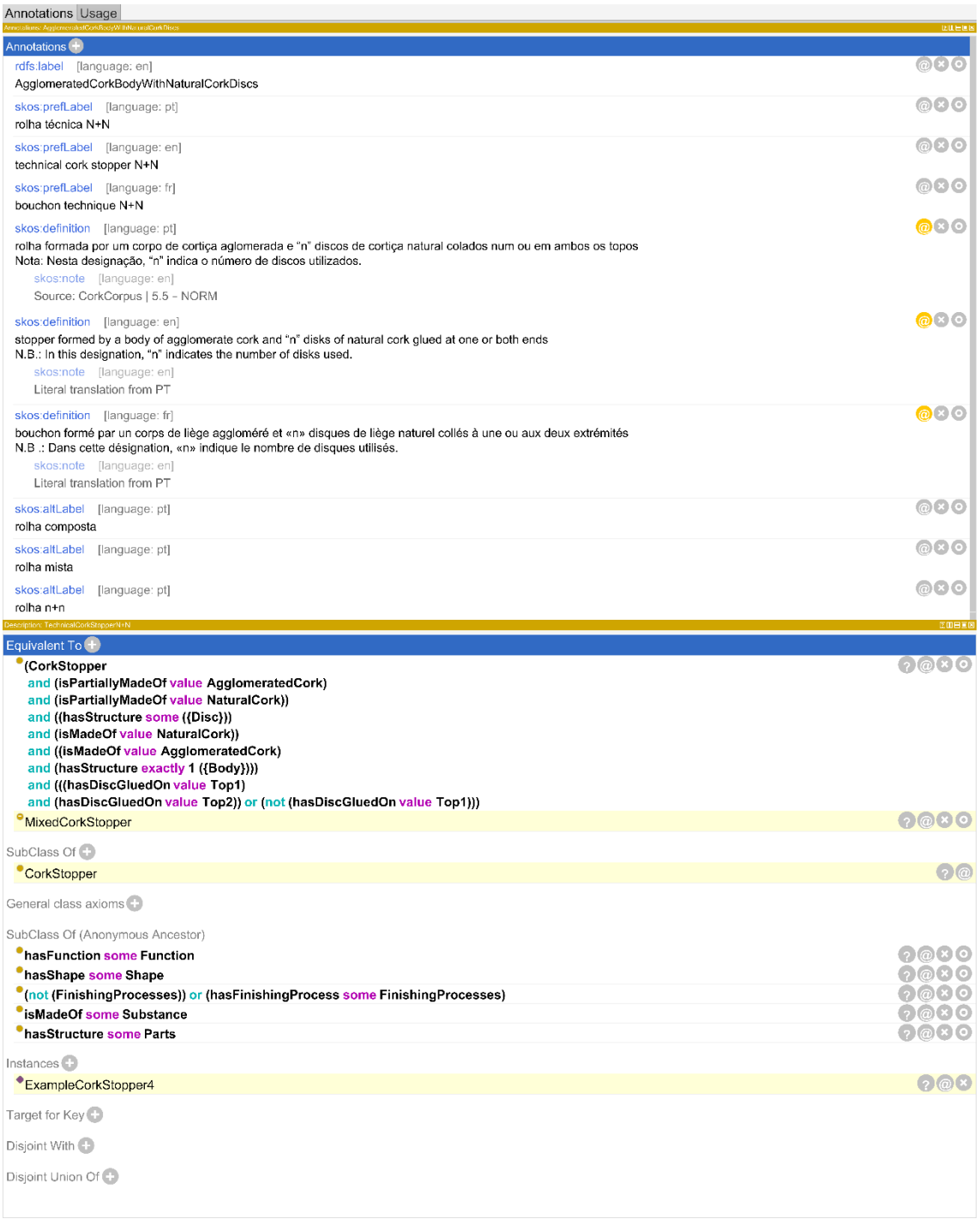


Figure 60: Description of <AgglomeratedCorkBodyWithNaturalCorkDiscs>.

Figure 60 above corresponds to the concept description of an `AgglomeratedCorkBodyWithNaturalCorkDiscs`.

As we can observe highlighted in yellow, an `AgglomeratedCorkBodyWithNaturalCorkDiscs` is a subordinate concept of `CorkStopper` and is classified as being an equivalent to `MixedCorkStopper`. The latter is disjoint with `NaturalCorkStopper` and `AgglomeratedCorkStopper`, and

therefore the reason why `AgglomeratedCorkBodyWithNaturalCorkDiscs` is not declared as disjoint with any other concept.

Furthermore, this kind of `CorkStopper` is composed of several `Parts`, namely 1 `Body` (made of agglomerated cork) and several `Discs`. The number of discs and the location where they are glued is directly related to the function of the cork stopper: cork stoppers for still wines or sparkling wines. Therefore, the different compositionality shown on the concepts' names.

This concept has 3 species, namely `AgglomeratedCorkBodyWithNaturalCorkDiscs1+1`; `AgglomeratedCorkBodyWithNaturalCorkDiscs2+2` and `AgglomeratedCorkBodyWithNaturalCorkDiscs0+2`.

We followed the same methodology as demonstrated in the linguistic and conceptual analysis of natural language definitions. Based on linguistic markers pointing at conceptual information, we have created conceptual relations identifiers to make explicit the intension of the concept, through the elicitation of all its characteristics in a formal definition.

As we can observe on the names of these concepts, the difference between `AgglomeratedCorkBodyWithNaturalCorkDiscs2+2` and `AgglomeratedCorkBodyWithNaturalCorkDiscs1+1` is numeric: one has 1 disc on each top whereas the other has 2 discs on each top. The same happens with `AgglomeratedCorkBodyWithNaturalCorkDiscs0+2`[178], yet, with a nuance: the number of discs is 2 but only glued on 1 top; the other top is empty=0.

The task of creating conceptual relations to make explicit all these differences was not straightforward, especially when it came to defining a concept with the same number of parts, but not in the same order, such as `AgglomeratedCorkBodyWithNaturalCorkDiscs1+1` and `AgglomeratedCorkBodyWithNaturalCorkDiscs0+2` (i.e., they both have 1 body and 2 discs). Instead of creating conceptual relations to describe this intricacy of parts –

[178] Out of curiosity, in specialised context discourse, the term of this concept is "technical cork stopper 0+2", whereas in non-specialised texts it is commonly called "champagne cork stopper".

which in our opinion would create an overwhelming set of definitional characteristics; a drawback since instead of a clear and unambiguous definition, the outcome would be crushingly noisy – we decided to elaborate descriptions restrained to these three concepts in the sense of additional information at the level of instances.

To this additional information, we will henceforth call descriptive characteristics.

The numeric difference and the location of the discs are the coordinates on which we based ourselves for the elaboration of the descriptive characteristics. For their elicitation, we used the owl:DataTypeProperty[179] – shown as Data property:

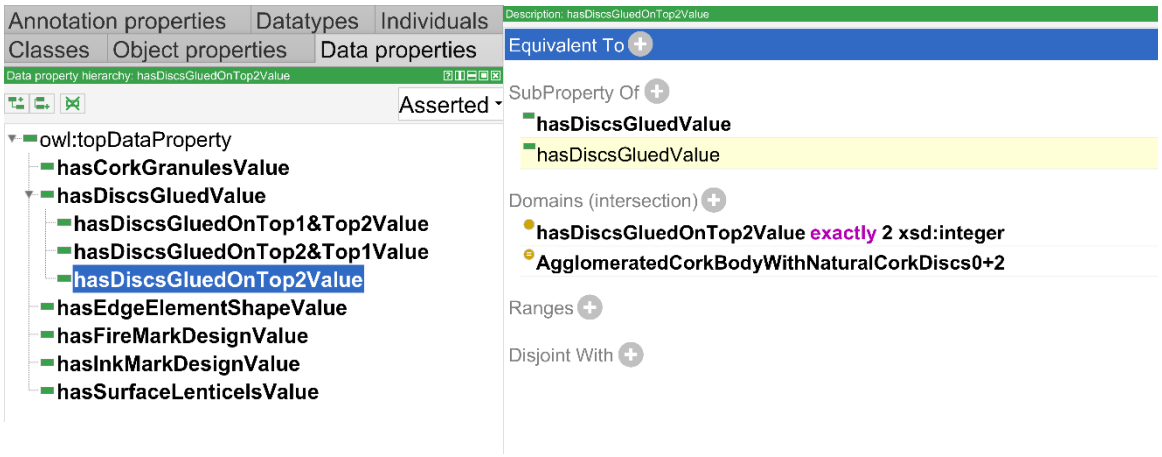


Figure 61: The data property hasDiscsGluedOnTop2Value and its domain, restricting the same data property to an integer: exactly 2, in addition to <AgglomeratedCorkBodyWithNaturalCorkDiscs0+2>

As represented on the right-hand side panel of Figure 61 above, the data property hasDiscsGluedOnTop2Value has its domain restricted to an integer: exactly **2**, in addition to `AgglomeratedCorkBodyWithNaturalCorkDiscs0+2`. With this descriptive characteristic, we can:

(i) differentiate an individual that has exactly 2 discs glued on only one top (Top2) from an individual that has "N" discs glued on both tops (Top2 and Top1); and

[179] This class "is a subclass of the rdf:Property class used to identify a property whose value is associated with a datatype. OWL datatype properties support "data-value" relations of instances. The values of "owl:DataTypeProperty" properties are literals […]. Datatype properties relate instances that belong to datatypes. The datatypes can be strings or simple XMLS datatypes." (Lacy, 2005, p. 170).

(ii) get a classification of its compositional type given the intersection of 2 domains, as shown below in Figure 62, (i.e., individuals described with such property can only be of the kind `AgglomeratedCorkBodyWithNaturalCorkDiscs0+2`), and finally

(iii) insert literal information, such as the name of the winery or the region of the wine's production – all the non-essential information but undoubtedly complementary to better understand or identify the concept in which the individual classifies for membership, as shown below:

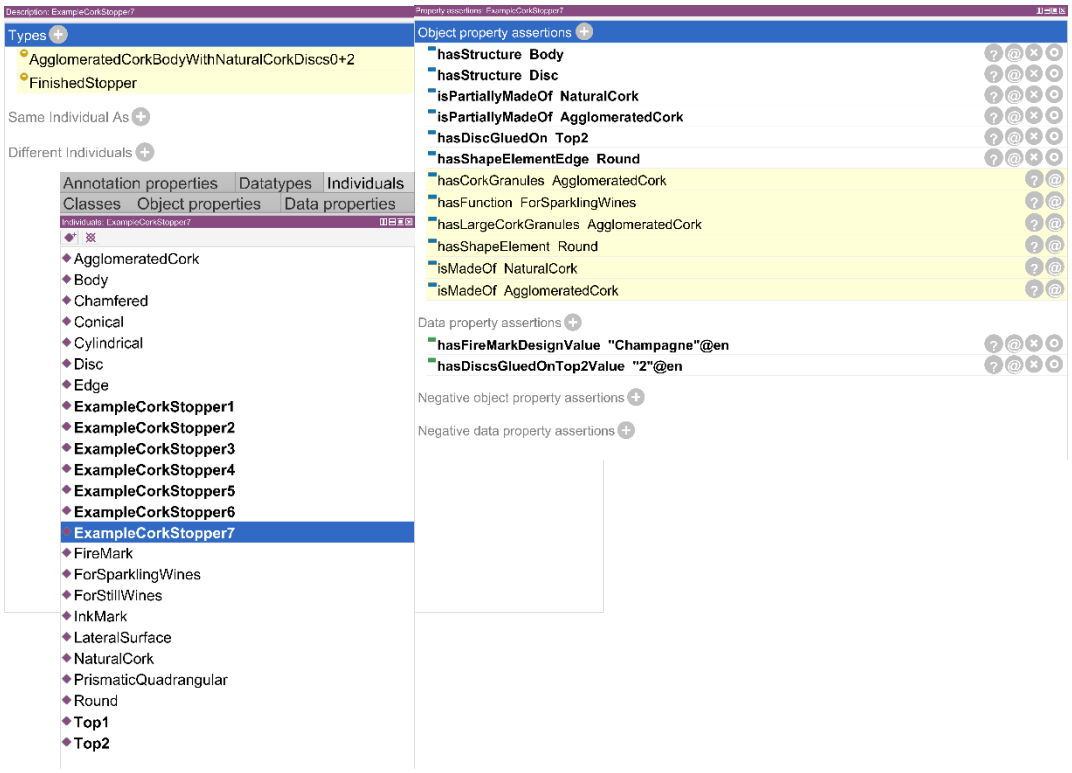


Figure 62: Classification of the individual ExampleCorkStopper7, as a kind of <AgglomeratedCorkBodyWithNaturalCorkDiscs 0+2>, as well as a kind of <FinishedStopper>.

On the left-hand side panel of Figure 62 above, we can see the reasoner highlighting in yellow that the individual ExampleCorkStopper7 (described on the right-hand side panel) is:

first, a kind of `AgglomeratedCorkBodyWithNaturalCorkDiscs0+2`,

and secondly, a kind of `FinishedStopper`.

The first classification is obtained from the domain property we have outlined above, and the second classification is obtained from the domain property restrained to the class `FinishedStopper` albeit with an additional restriction: the range set to literal[180] – where any string of characters may be inserted.

6.6. Some remarks: the long name of concepts

As a final word, we would like to clarify the reason why the names of the concepts in our ontology are so long, such as <InkMarkedColmatedWashedMonoPieceNaturalCorkStopper>, and offer some concluding notes.

The name we have chosen to label this concept is rather wordy, like for most of the concepts in this ontology. There is a deliberate purpose underlying such verbose names: concepts are named with a label according to their characteristics, not just with the differential characteristics but with the whole set of characteristics that compose the concept. The idea is to visualise the *intension*[181] of the concept. Consequently, the more specific concepts get through the addition of differential characteristics, the longer their name becomes. An advantage of this methodology is that one can rapidly have access to the information being pointed at by the components of those verbose labels, as opposed to what occurs in specialised discourse contexts. We have observed in specialised texts that terms – the verbal expressions of concepts – are commonly expressed in a reduced form. In our opinion, this is a consequence of the expert's knowledge since he/she knows the intension of the concept and therefore has no need to make the underlying knowledge explicit. As an example, an `InkMarkedColmatedWashedMonoPieceNaturalCorkStopper` is uttered as "marked stopper": a polylexical unit, whose morphosyntactic structure has only one

[180] Keeping in mind that a property is a binary relation, datatype properties are different from object properties: they establish a relation between instances of classes and RDF literals and XML Schema datatypes (see W3C, 2014).

[181] According to ISO 1087-1, it is the "set of characteristics that make up a concept" (2000, p. 3).

linguistic element pointing at an operation, which is "marked". This operation is the very last one in the manufacturing process, regardless of all the others that have occurred before. The same applies to most terms designating `CorkStoppers` submitted to operation(s).

Looking further at "marked stopper", we can observe that experts tend to name concepts according to their differential characteristic(s), from an Aristotelian definitional point of view; a tendency that is useful to perceive conceptual information, but, if, and only if, there is access to the unspoken information. In our view, this is where an ontology plays its fundamental role when it comes to unambiguous communication.

Access to that unspoken information can be achieved through the ontology, where the conceptual system of the domain is explicitly represented by means of formal definitions. However, we must stress that this does not occur exclusively through these definitions. We believe that this methodology – naming concepts with long names, in which the concept's *intension* is explicit – is an added value that enables non-expert users to rapidly perceive what is the location of a given concept in the *concept system*[182] of the domain and immediately understand what type of substance it is, its shape, the number of treatments it has undergone, and so forth, by virtue of the concepts' name.

Furthermore, for the terminologist-linguist, such long names are almost a requirement for the creation of the ontology: not only the place of the concept is far clearer in relation to its neighbour concepts in the concept system, but it also prevents hierarchically misplacing the concept in the ontology: a `InkMarkedColmatedWashedMonoPieceNaturalCorkStopper` is clearly a specification of a `ColmatedWashedMonoPieceNaturalCorkStopper`.

Taking into consideration what has been said, the intension of the concept explicitly stated in its name – rdfs:label in the ontology – is first and foremost seen as a complementary guideline to writing textual definitions, considering that all the characteristics of the concept are stated. This leads to a question: What model should we follow when writing textual definitions? Should we mention every single

[182] According to ISO 1087-1, it is the "set of concepts structured in one or more related domains according to the concept relations among its concepts." (2000, p. 6).

characteristic in the textual definition, thus creating a definition pointing at as much information as possible? Or should we stick to the classical Aristotelian model, in which the intension of the genus is implicit by heritage, and therefore the differential characteristic(s) are the unique additional information to distinguish the concept being defined from its neighbouring concepts – parent or sibling? The answer does not seem straightforward; thus, another question should be asked to help us answer the previous two: To whom and to what purpose are we writing the textual definition?

Without deviating from our terminological perspective, and bearing in mind that Terminology is a science that studies the terminology (i.e., the set of terms) of a given field of expertise through terminological work (the systematisation of concepts and corresponding terms) so as to build terminology resources, such as glossaries and knowledge databases, as outcomes from terminological data retrieved from corpora, such as definitions, contexts and terms – verbal designations of concepts – the methodology for writing definitions in this context must likewise follow the epistemological principals of this science. Thus, a textual definition – a text through which the *description*[183] of a given concept is stated in natural language – should convey essential and well-structured information, such that a conceptual micro-system is represented; a path through which the position of that concept is inferred. Well-structured information implies a description of the concept where all differential characteristics are explicit as far as a hierarchy is represented by the inclusion of the *proximum genus* – a concept that is (should be) already defined in our terminological work. Such descriptions are called *terminological definitions* (Rey, 1979; Pavel & Nolet, 2002) or *definition by intension*[184] (Felber, 1984); a concept description to which we have resorted throughout this study in the terminology of ISO: ***intensional definition.***

[183] According to Felber, "A definition is a **description** of a concept by means of other concepts, mostly in form of words and terms. It determines the position of this concept in a system of other related concepts." (emphasis added)(1984, p. 160).

[184] "consists of a specification of the characteristics of the concept to be defined, i.e. the description of the intension of the concept." (Felber, 1984, p. 160).

Hence, one of the core tasks in specialised information management is to differentiate what is essential information from what is inessential to be included in the text of a natural language definition.

The criteria for the decision of maintaining or discarding essential and inessential information is a task that directly depends on the knowledge of the domain under analysis. As far as we could understand, `Washing` is an operation that is always present in the manufacturing process of cork stoppers, or duplicated, even triplicated, throughout the process. If this operation is withdrawn from the description of the concept, the concept remains identical, except for the description of a clean state. On the contrary, the characteristic `Shape` is what defines and/or differentiates a given `CorkStopper` from the ones that remain in a cylindrical plane form. The same can be asserted regarding its compositionality: Is a stopper composed of one or several parts? The relevance of compositionality, in the case of `NaturalCorkStoppers`, is quite significant, for no other type of natural cork but "Amadia" is suitable for the mono-piece manufacture.

In this case, we would suggest that `Washing` is a piece of inessential information to write a natural text definition, although being significant at the level of a formal definition for the purpose of being classified as `Semi-finishedCorkStopper`. Thus, looking at the name of the concept `InkMarkedWashedColmatedWashedMonoPieceNaturalCorkStopper`, the element `Washing` is one of the characteristics present in the concept name that can be elided from the text of the natural definition, when occurring duplicated or more.

We have systematised this reflexion below, in Table 30, in which the concept `WashedMonoPieceNaturalCorkStopper` gets more specific from the bottom to the top (from 3 to 1).

Table 30: *The complementarity of the linguistic and conceptual information*

1	**concept name**	***Ink marked***	***colmated***	***washed***	***mono-piece***	***natural***	***cork***	***stopper***
	differential characteristic	characteristic						
	proximum genus		characteristic	inessential characteristic	characteristic	characteristic	characteristic	genus
	Transcription into the Aristotelian formula							
	X = Y + DC		colmated	mono-piece	natural	cork	stopper	ink marked
	X (specific concept) =		**Y (proximum genus) +**					**DC (differential characteristic)**
	Textual definition proposal							
	def= colmated mono-piece natural cork stopper that was ink marked							
2	**concept name**		***Colmated***	***washed***	***mono-piece***	***natural***	***cork***	***stopper***
	differential characteristic		characteristic					
	proximum genus			inessential characteristic	characteristic	characteristic	characteristic	genus
	Transcription into the Aristotelian formula							
	X = Y + DC			mono-piece	natural	cork	stopper	colmated
	X (specific concept) =			**Y (proximum genus) +**				**DC (differential characteristic)**
	Textual definition proposal							
	def= mono-piece natural cork stopper that was colmated							
3	**concept name**			***Washed***	***mono-piece***	***natural***	***cork***	***stopper***
	differential characteristic			inessential characteristic	characteristic			
	proximum genus					characteristic	characteristic	genus
	Transcription into the Aristotelian formula							
	X = Y + DC				natural	cork	stopper	mono-piece

X (specific concept) =	Y (proximum genus) +	DC (differential characteristic)
Textual definition proposal		
def= natural cork stopper entirely made of “Amadia” cork		

Based on the Aristotelian formula X = Y + DC, we have separated all the elements that constitute the concept's name, each of those representing one characteristic, as shown in Table 30 above.

Considering that X = is the concept being defined, Y = the intension of the superordinate concept and DC = the differential characteristics, all of the recorded characteristics along X are taken into consideration to write a natural language definition (def=), except for the information considered inessential for this task, namely "washing". Given our knowledge of the domain under analysis, we have correlated the characteristic /MonoPiece/ to the concept `AmadiaCork`, as we can see for the definition (def=) of the concept 3 in Table 30.

On the other hand, `AmadiaCork` will necessarily be defined in our ontology. As a common practice in terminological work, all concepts referred to are defined. If we had written a definition for concept 3 like, for instance, *def=a natural cork stopper composed of one piece*, it would not be totally wrong. Similarly to definitions (1) and (2), we would be replicating the exact same information that it is already being conveyed by the name of the concept. However, a sense of incompleteness would prevail, if there was not a reference to the information `AmadiaCork`.

We would like to emphasise the complementary of the approach exemplified in Table 30 above, particularly concept 3. Their first feature is the ability to facilitate the task of writing intensional natural language definitions given their linguistic label (i.e., by reflecting the amount and order of operations involved in the manufacturing of the stopper). However, we must stress that natural language definitions cannot be automatically generated, nor are formal definitions – in the sense of ontology editors – sufficiently human-readable so as to provide comprehensive textual definitions for communication purposes. In our opinion, these formal and non-formal methods are complementary mechanisms. When the terminologist-linguist is building an ontology with formal definitions systematically based on those long concept names mentioned above, she/he can also systematically create intensional definitions in a coherent and logical manner. Finally, when the intension of a given concept becomes too generic, the corresponding natural language definition might incur in incompleteness, as demonstrated with definition 3 in Table 30 above. It is here where the

management of terminological information plays a complementary role, in the sense of either replacing or inserting complementary information, such as /MonoPiece/ + /natural/, which corresponds to `AmadiaCork`.

6.7. Some conclusions

As a part of our terminological work, the task of systematising concepts resorting to the theory of sets and to subsumed classes, along with logic principals, has proven to be a valuable advantage for the pursuit of writing coherent and well-formed intensional definitions in natural language. We have come to this conclusion given the tasks of organising concepts in a systematised way, resorting to logical constructs – in the sense of DL. The multifaceted methodology we had to develop in order to put those tasks in practice, permitted us to confirm that terminology is a multidisciplinary scientific field of knowledge.

This multidimensional perspective is not recent within the community of terminologists (see Rey, 1979); on the contrary, given the ultimate goal of all those spheres of knowledge we have mentioned above, the study of the concept has been substantiated for a while. This is where we have to ask: How do concepts relate to the theory of *logical classes*[185]? According to Rey (1979), despite the closer interconnection of logical classes to the mathematic set theory given the major focus of the former on formally representing membership of abstract objects to sets, Terminology – as well as semantics and non-formal logic – should also look for sets of common traits, albeit for concept organisation purposes. This means that common features are what underpins class membership, which, in turn, assists in the organisation of concepts in a system. Such approach is called semantics in

[185] In the sense of Rey: "L'apparition d'universaux logiques repérés au moyen des noms est liée à la constitution de classes par lesquelles les spécificités individuelles des « particuliers », des objets concrets distingués par l'esprit – eux-mêmes constitués par abstraction généralisantes à partir de suites d'information (perceptions, etc.), réunies elles aussi en classes – sont écartées ; d'autre caractères, communs et hiérarchisables, étant seuls retenus.[…] Le pouvoir désignatif, dénominatif, d'un signe (mot, nom ou terme) se ramène à la classe des référents auxquels il correspond. D'un point de vue onomasiologique (désignation), les classes référentielles sont fondatrices du concept." (1979, p. 35).

comprehension or intensional semantics (in opposition to extensional), where the fundamentals of the art of creating a *definition* can be found.

In our opinion, this is the bridging point where Terminology and the Theory of classes shake hands. Our reflexion derives from the assumption that classification is indispensable both in the scope of knowledge spheres and in the empirical daily experience, where stable designations are formed (see Rey, 1979).

According to Rey (1979), terminologies are systems of names and definitional systems, for they correspond to the only concrete realisation – in the form of signs of a language – of notional systems. Compared to the real world, the constructive feature of definitions, and subsequent concepts they express, depends on a categorisation – an organisation by means of logic universals (as opposed to individuals) – therefore, providing the possibility of an intensional analysis and comprehending the characteristics of the concept. The organisation of such definitions entails a structure. It is this structure that withstands a domain that rightfully assigns to the names belonging to these systems the status of authentic terms.

In sum, the act of defining is closely related to classification inasmuch as knowledge is perceived through the set of interwoven characteristics. The set of characteristics is the definition proper, corresponding thus to a concept logically related to other concepts given its own defining characteristics. The ontology we have built corresponds to a structure of intensional definitions thoroughly organised to the extent that a domain concept system is coherently mirrored. The concepts that form this system underpin the terminology of the domain under study, for the *referential*[186] *meaning* they carry consistently denotes the entity expressed by the term.

[186] In the sense of Kleiber: "le sens obéit à deux modèles référentiels différents : le modèle descriptif, celui qui indique quelles sont les conditions (nécessaires e suffisantes ou prototypiques) auxquelles doit satisfaire une entité pour pouvoir être désignée ainsi, et le modèle instructionnel, qui marque le moyen d'accéder au, ou de construire le référent. Le premier est prédicatif, le deuxième met en jeu des mécanismes dynamiques (déictiques, inférentiels), qui ne constituent pas des propriétés du référent, mais des balises plus au moins rigides pour y arriver." (1997, pp. 32-33).

CONCLUSION

Overall remarks

This thesis derives from an academic work where we demonstrated that linguists can infer the conceptualisations of experts from texts. Furthermore, we have proven that the automatic processing of texts is not enough unless we complement it with a rigorous linguistic analysis of texts by examining the terminological choices made by the experts and uncovering the underlying lexical-semantic relationships in order to infer specialised knowledge information.

With this research, we intended to prove that it is definitely possible to identify concept by analysing texts. The key is not to confuse the linguistic and the conceptual levels of analysis and look at the data beyond their automatic extraction. Texts are undoubtedly vehicles of knowledge transfer. Analysing texts to extract the concepts' linguistically expressed characteristics effectively allowed us to grasp conceptual relations that are specific to the domain. As demonstrated in this study, we were able to propose a preliminary conceptual organisation of the subject field. Although not addressed in this study, the outcomes of this study are worthy of discussion with experts in some future work.

On the other hand, we used Protégé and concluded that it can be a useful tool for linguists. In our view, Protégé is an interesting tool, mainly because it allows us the complementarity of both the conceptual and the linguistic dimensions, and also from the metadata point of view (essential for interoperability in the semantic web and LOD[187] context). The personal challenge, which we found very stimulating, was to overcome the limitations of the tool and use it effectively to achieve our goals.

We have thus bridged three main aspects in our study: (i) the classical aspects of the Aristotelian logic; (ii) the methodology of our terminological work – where characteristics play a fundamental role in the analysis or the elaboration of intensional definitions; and finally (iii) formal definitions, for which we resorted to Protégé and inherent *Web Ontology Language*

[187] https://lod-cloud.net/

(OWL) – a W3C Recommendation – to formally describe the concepts of the domain to relate them via high-level abstract syntaxes and formal reasoning in a reason-*able* ontology once concepts are consistently defined.

Some insights

The analysis of the specialised corpus we compiled and particularly of the definitions we were able to extract from it has allowed us to make the following considerations.

Natural language definitions are terminological information-rich contexts and are therefore excellent sources of specialised information to achieve our purpose of making a terminological study on the domain of cork. In the present study, we have achieved an organisation of concepts in this specialised domain by systematising the terms we extracted from texts. We conclude that the methodology proposed in this study has positively responded to our goals, despite being extremely time-consuming – an inevitable consequence from the multidisciplinary tasks of this terminological work:

Throughout our analysis, we were able to see that there are different types of definitions: (i) intensional definitions − those that point to the intension of the concept by indicating the superordinate concept and the distinctive characteristics; and (ii) contextual definitions − those that describe the concept by using textual segments that explain "what the thing is".

The analysis of the different types of definitions allowed us to identify different types of information with the purpose of finding out how to take advantage of existing definitions to extract semantic and terminological information.

Texts are the linguistic expressions of knowledge, which makes them an inescapable subject for terminological work. The double dimension of Terminology is thus ensured since the description of the concept is not confused with the concept itself. The linguistic markers extracted from the defining statements have proven to be very productive for the creation of lexical maps, which have allowed us to identify, for example, Aristotelian definitions that comply with the following formula: X=Y + DC.

Linguistic markers (LM) figure prominently in our study. The lexical and semantic relationships that LM establish between two terms or between other forms and their

characteristics allow us to organise linguistic knowledge, i.e., they enable us to determine the essential and distinctive characteristics of the concept defined in natural language, arranging terms in lexical networks and subsequently suggesting conceptual maps that underlie the elicitation.

Specialised corpus builders and particularly terminologists that decide to create a domain specific corpus from scratch must keep in mind that the first task is to become familiarised with the subject field. The task of familiarisation is of utmost importance, not only for the corpus compilation task since it is essential that we ascertain whether a specialised text is authoritative, but also for the interpretation of the corpus outcomes, i.e., the textual data extracted, either observed in the concordances of a word of interest (KWIC) or in a given piece of information that is not complete. As we have highlighted, some specialised texts do not always convey explicit information.

Regarding the process of building a corpus from scratch, we should emphasise that corpus design is what establishes *a priori* what type of corpus is going to be constructed; thus, a set of well-defined criteria ought to answer the corpus purposes. Criteria are not identical for each and every corpus since each corpus-driven work is different. Thus, the criteria suggested in the corpus linguistic literature are not mandatory, nor finite in number. Each research has different goals; therefore, different corpus designs are possible. Furthermore, once the set of criteria is designed, it is essential that corpus builders have the perception that sometimes, the rigidity of the criteria may introduce some difficulties, such as the date of publication, for instance. In that particular case, exceptions must be thought of considering that older texts may contain information that modern texts do not. The same can apply to any other criterion if rigorous constraints are not at the core of the corpus' purposes. The key is to record every decision made during the corpus compilation stage, so that future users (if that is the case) will effectively interpret the corpus outcomes.

Since our aim was to deal with specialised discourse, we defined as an essential criterion for the compilation of our corpus that texts produced by experts would be included, given that these are governed by the pragmatic constraints of the context: the higher the expertise of the interlocutors, the more specialised the discourse. Hence, the communicative setting is a core-criterion for a domain-specific corpus compilation, although not exclusively.

The rationale behind this criterion is the fact that texts produced in such communicative setting are rich in definitional contexts and/or contexts that describe concepts given the different degrees of knowledge of producers and recipients. The expectation underlying this predefined criterion was achieved in our study: we have created a text collection that coherently represents the social discourse community of the professional, technical, and scientific practice within the special field of cork.

The compilation and exploration of the corpus were both performed with the Sketch Engine (SKE) software. SKE is a complete and powerful tool that allows corpus builders and users to use a series of built-in features, such as, for instance, automatic corpus annotation – albeit POS tags only – and advanced corpus search, i.e., using a corpus query language (CQL) in which regular expressions (regex) resorting to python symbols are creatively used. Natural language processing (NLP) tools have limitations, and SKE is no different. The results of the built-in feature Sketch word that we compared with the results obtained by regex have proven that the FreeLing tagger cannot distinguish between adjectives and past participles, which is highly limiting as regards terminological work. However, we have demonstrated that the possibility of exploring the corpus with text mining strategies, i.e., using regex, is a definite plus: depending on the user's skills and creativity, the results can be even more tuned to the user's expectations. The highlight here is that text mining strategies do not straightforwardly provide us with optimum results. Instead, it is an iterative work that is time-consuming for it is necessary to create a new regex after the expectable or silent results from the previous regex. Finally, each of those results must first be linguistically and then conceptually analysed, for knowledge extraction is our ultimate goal. For us, statistics measurements of a given linguistic expression are a complementary feature that can corroborate (or not) that we have found a linguistic pattern.

The structured organisation of the domain's terminology is the terminological work itself. This assumption explains the close link between term and definition. In Terminology, the intensional definition plays a fundamental role, for the structure of the information made explicit in the intensional definition leads to the recognition and differentiation of a given concept. As pointed out by ISO 704 (2009), intensional definitions are considered the most explicit and precise method of concept definition. Thus, for the task of writing such definitions,

terminologists must focus on the characteristics that constitute a given concept, for the set of characteristics is what defines the concept and/or differentiates it from other concepts. Essential characteristics are fundamental to define a given concept. However, although not fundamental, descriptive characteristics play a substantial role in the formal knowledge organisation descriptions, for they provide us with the means to make explicit information at the level of individuals, as shown for the shift of status (i.e., from <Semi-finishedStopper> to <FinishedStopper>) in the ontology.

Finally, there is no ideal model to create an ontology beyond a list of good practices that must be taken into account, such as coherently follow the chosen epistemological perspective of what a concept is and the formal language (description logic) that assists the ontology editor, such as Protégé. The objectives of our ontology seek to respond to two typologies: (1) the type of cork stopper compared to the type of cork (raw material) with which it is produced; (2) the typology of operations that belong to the finishing processes. Finally, this ontology should also respond to the state of completion – in the sense of finished product – of the cork stopper, depending on the last operation to which it was submitted.

Building an ontology that addresses those three aspects led us to hesitate between two models, which in turn made us question:

What is the best criterion for modelling the concepts that represent objects submitted to processes, where several operations (activities) occur?

(a) Through the subsumption of a given <Cork Stopper>, whose systematisation follows the differential division of the substance (e.g., <Natural cork> vs. <Agglomerated cork>) and each concept is described by the accumulation of processes, where for each process involved, a new concept is subsumed. Here, we obtain an extension of a given genus through the differential characteristics (i.e., a process) introduced in each new concept; or

(b) Through the subsumption of a given <Cork stopper>, whose systematisation also follows the differential division of the substance, but where all concepts are subsumed by the same genus, despite being described with every operation that has occurred. The result is that each such concept is the most specific concept of the genus extension

(e.g., <Natural cork stopper> is the genus of all <Natural cork stoppers> + <Operation>). Thus, each of these concepts establishes a horizontal relation (a sibling relation).

Similar to option (a), the concept <FinishedStopper> is a specification of <Semi-finishedStopper. Nevertheless, with the systematisation stated in option (b), we obtain, on the one hand, a distinct separation between the concepts that are classified as <FinishedStopper> from those that are classified as <Semi-finishedStopper> when questioning the ontology; however, on the other hand, it does not satisfy us at the level of the extension of a concept that represents an object that has undergone several operations. Thus, the notion of process is not conveyed, as demonstrated below in Figure 63:

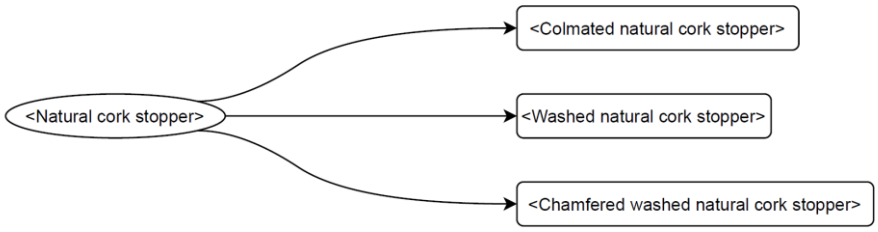


Figure 63: Representation of the systematisation of concepts according to option (b)

According to Figure 63, a <Washed natural cork stopper> and a <Chamfered washed natural cork stopper> are siblings. This is correct when we consider that they are both concepts that classify as <Semi-finished stoppers>; however, these concepts belong to different levels of specification, for we have the acknowledgment that the first two concepts denote only one <Operation>, while the third denotes two <Operations>, as demonstrated by the long name of the concept in Figure 63.

Furthermore, in our view, option (b) brings us closer to natural language, namely the discursive level. At this level, the information is not totally verbalised since the intersubjectivity of the experts allows them to understand the missing information, e.g., despite being submitted to the operation <Washing>, the term that designates <Chamfered washed natural cork stopper> is “chamfered stopper” – the last operation is the one the

expert verbalises in discourse. Consequently, a sense of a unique <Operation> is ambiguously conveyed.

Contrastively, option (a) allows us to represent the unspoken information. The concept of <Stopper> is specified through the accumulation of <Operations> (i.e., characteristics). It therefore conveys more information than the concept described by the last <Operation> to which it was submitted, and consistently mirrors the flow of a process, as shown below in Figure 64:

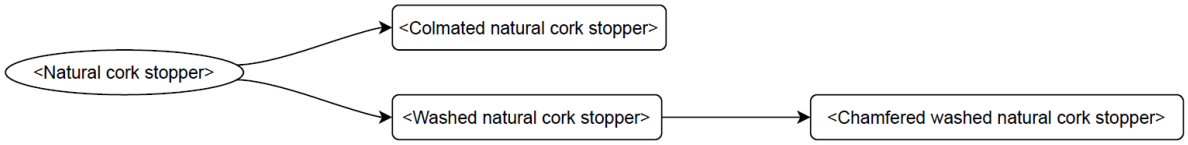


Figure 64: Representation of the systematisation of concepts according to option (a)

These insights led us to choose the model stated in (a) for it is on the set of all operations, i.e., additional characteristics to the *proximum* genus concept, that we rely to propose a model for writing intensional definitions.

In sum, this study intends to lay a methodology for the analysis of definitions written in natural language. It is an iterative work involving several tasks and several steps, where each of these depends (or not) from the previous one, both in the linguistic and the conceptual dimensions. The gold standard of this analysis is, first and foremost, to interpret the meaning to which each term is pointing at, along with the meaning of the lexical marker that intermediates the establishment of a lexical-semantic relation between those terms so that one can infer special knowledge information. This information is conveyed by the types of lexical relations expressed by those lexical markers and the terms they relate in the syntagmatic axis. Finally, it is the interpretation of these lexical relations that allows us to obtain conceptual information not in a straightforward manner but through several mechanisms, for the two dimensions of Terminology – conceptual and linguistic – are not isomorphic.

Future work

The primary goal of this terminological work is creating a terminological database to feed a terminological resource that gathers several levels of information, namely linguistic (plurilingual), conceptual and visual. Congregating several semiotic levels of information is what defines our project as a multi-semiotic terminological resource.

Following the goals of this study, the purpose of building a domain-ontology is primarily to organise concepts denoting the real-world object designated by the term "cork stopper", in a systematic way, depending on the type of substance the object is made of, its function, shape and parts. The substance the object is made of plays a central role: concepts are hierarchically systematised according to their composition, in the sense of substance (raw material). The ultimate purpose of the ontology is to organise the operations intervening in the manufacturing process of the objects – the "cork stoppers" – depending on the purpose of each operation in order to obtain a classification of each object. Finally, this classification aims at providing information on the manufacturing stage of the object regarding its stage of completion within the manufacturing process. As a whole, the ontology aims at providing (1) a typology of objects and (2) a typology of operations, connected through an intricacy of logical relations. From here we obtain a classification of the type of object, regarding its properties and stage of completion given their interdependency.

The applicability of such ontology is multifarious. It can be a valuable asset for language professionals – such as translators, for instance – as well as for the industry. As an example, the ontology here designed could be of assistance in the development of a model to monitor the phases of a given manufacturing process.

The ultimate goal of the TermCork project is creating an e-dictionary designed as a multilingual and multimodal product, where several resources, namely linguistic, conceptual, and multimedia, are paired to facilitate knowledge acquisition by the user. It is here where the collection of images and videos we have captured along the corpus compilation stage plays its role, as well as the publishing dictionary editor Lexonomy; hence, the notion of multimodality. In order to accomplish this goal, we intend to link the several resources we have developed, namely both the CorkCorpus and OntoCork to Lexonomy, as depicted below.

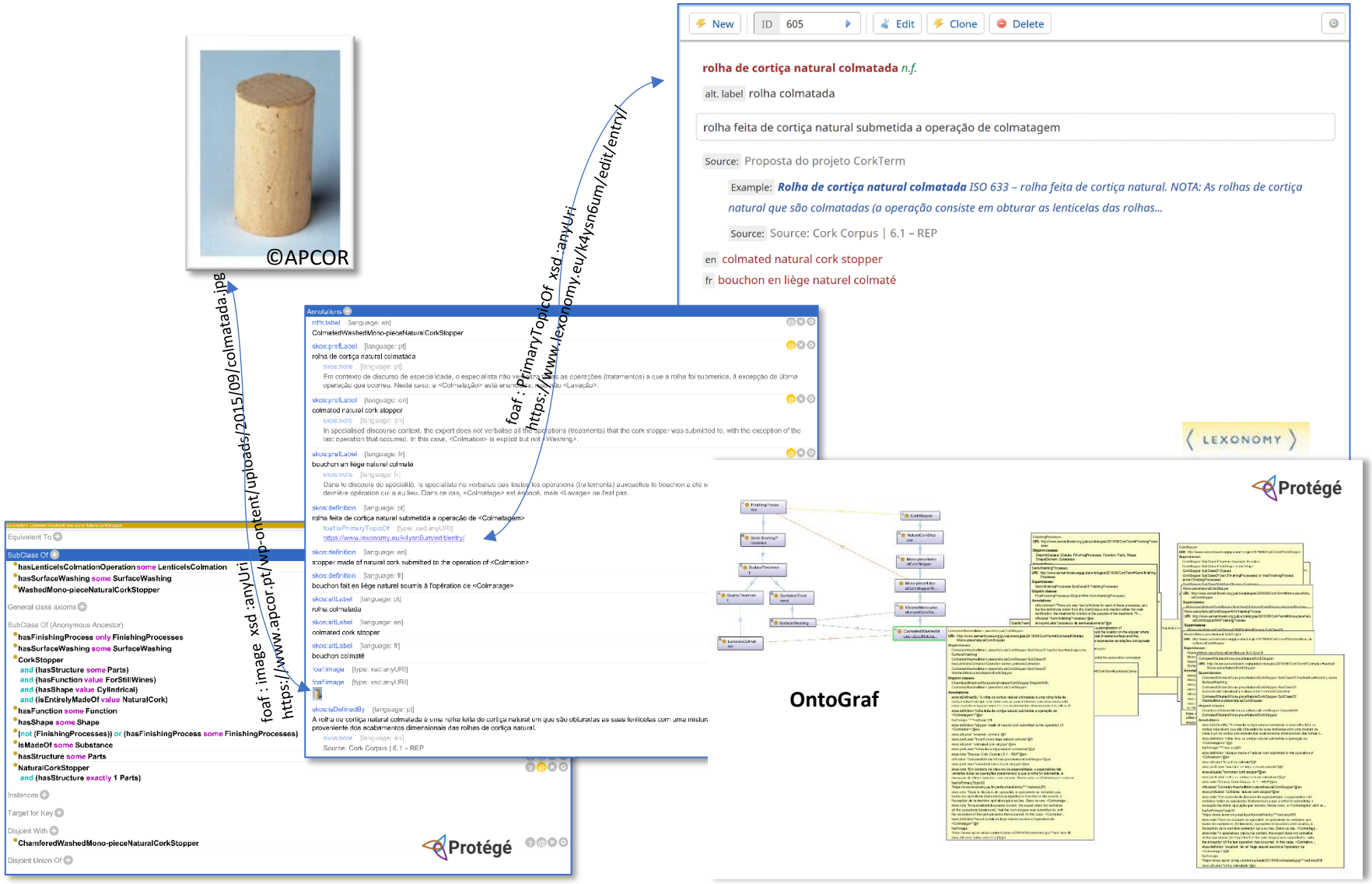


Figure 65: Representation of the formal description of <ColmatedWashedMono-pieceNaturalStopper> linked to two resources: an image and a dictionary publisher.

As shown above in the interface of Lexonomy, on the top right-hand side of Figure 65, we have elaborated an intensional definition for the concept <ColmatedWashedMono-pieceNaturalCorkStopper>, written in Portuguese, and followed below by an example written in blue. This example was drawn from the Cork Corpus, more specifically, from a text type report labelled with the descriptor "6.1 – REP", as shown in the interface after the label "Source". The English and French equivalents are included at the bottom of the interface.

The interface Lexomomy can be accessed from Protégé, through the annotation we have inserted in the skos:definition resorting to the FOAF Vocabulary Specification 0.99[188]. However, this access does not directly open the definition we are pointing at, but to the whole set of definitions nested in the interface – an aspect that deserves future study.

An image was also added to this skos:definition, resorting to the same vocabulary. The FOAF project uses W3C's RDF technology, therefore the possibility of being used in Protégé, for underlying this RDF technology, "things" of the internet are linked by their Uniform Resource Identifiers (URI)[189].

In Lexonomy, the entry of the definition is the term "rolha de cortiça natural colmatada", whose grammatical category is a feminine noun, as noted under the descriptor "*n.f.".* This term is what we consider the preferred term. Under the preferred term, there is an alternative form: "rolha colmatada" positioned in line with the descriptor "alt. label". This last descriptor is based on the skos labels we have used on the Protégé editor. The purpose of this alternative form is disclosing another descriptor of the concept, in this particular case, identical to one of the forms found in the corpus. As mentioned before, this form is a term that contains reduced information but is commonly used in the specialised context of communication. Thus, if searched in the resource, the definition of the concept is obtained through this alternative form.

[188] For more details see Brickley & Miller (2014).

[189] According to the W3C organisation, "(URIs, aka URLs) are short strings that identify resources in the web: documents, images, downloadable files, services, electronic mailboxes, and other resources. They make resources available under a variety of naming schemes and access methods such as HTTP, FTP, and Internet mail addressable in the same simple way." (W3C, 1997).

To conclude, we consider we have demonstrated that the above methodology is an added value for the task of writing definitions in natural language. In our view, when defining a concept of a given domain in natural language, one must follow the formula of an intensional definition as recommended by ISO 704 (2009). The domain under analysis here is no different. Similarly, the formalisation of definitions employing a formal language should also follow this formula, where a DC is added to the *proximum* genus for each specialisation. In doing so, the actual observation of the concept formally described through logic constructs will support the coherency of its definition in natural language. The same applies to the representations of the knowledge defined informally in the first stage with CmapTools, and subsequently logically represented with an OntoGraf in the ontology. The actual observation of these knowledge representations is an added valued for the task of defining a concept since the consistency of definitions is complemented; thus, sharpening one's critical sense over the definitional texts from which we have started.

BIBLIOGRAPHY

Agbago, A., & Barrière, C. (2005). Corpus Construction for Terminology. *Corpus Linguistics 2005 Conference.* Birmingham: National Research Council of Canada.

AICEP. (2014). aicep Portugal Global. (aicep, Ed.) Lisboa. Retrieved 12 07, 2014, from aiecp Portugal Global: http://www.portugalglobal.pt/

APCOR. (2010). Rolhas de cortiça. *Cortiça. Cultura, Natureza, Futuro.* Cork Information Bureau 2010.

APCOR. (2011). Manual Técnico - Rolhas. Santa Maria de Lamas: APCOR.

APCOR. (2014). Rolhas de Cortiça. Santa Maria de Lamas: APCOR.

APCOR. (2015). *Cortiça - Estudo de caraterização setorial.* Santa Maria de Lamas: APCOR. Retrieved from https://www.apcor.pt/portfolio-posts/estudo-caracterizacao-sectorial-e-prospectivo-2015/

APCOR. (2019). *Anuário de cortiça 18/19.* APCOR.

APCOR. (2019). História. Information Bureau. Retrieved from APCOR: https://www.apcor.pt/media-center/press/

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing, 7*(1), 1-16. doi:10.1093/llc/7.1.1

Baader, F., & Nutt, W. (2003). Basic Description Logics. In F. Baader, D. L. McGuinness, D. Nardi, & P. F. Patel-Schneider (Eds.), *The Description Logic Handbook: Theory, implementation, and applications* (2nd Edition ed., pp. 47-100). Cambridge University Press.

Baader, F., Horrocks, I., & Sattler, U. (2009). Description Logics. In S. Staab, & R. Studer (Eds.), *Handbook on Ontologies* (2 nd ed., pp. 21-43). Dordrecht; Heidelberg; London; New York: Springer.

Baader, F., Horrocks, I., Lutz, C., & Sattler, U. (2017). *An introduction to Description Logic.* Cambridge: Cambridge University Press.

Baker, P., Hardie, A., & McEnery, T. (2006). *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Barata, & Ganhão. (2004). Caracterização Processual e Económico-Financeira do Subsector Transformador e Comercial das Rolhas de Cortiça Naturais e Aglomeradas. Lisboa.

Barrière, C. (2004). Building a concept hierarchy from corpus abalysis. doi:DOI: 10.1075/term.10.2.05bar

Bernauer, J. (1994). Subsumption Principles Underlying Medical Concept Systems and their Formal Reconstruction. *NCBI - National Center for Biotechnology Information*, 140-144. Retrieved from https://www.researchgate.net/publication/15249146_Subsumption_principles_underlying_medical_concept_systems_and_their_formal_reconstruction

Bicho, M. (2004). *A Rolha de Cortiça: da floresta à utilização.* Santa Maria de Lamas: APCOR-Associação Portuguesa da Cortiça.

Boshmonart, J. R. (2011). Environmental evaluation of the cork sector in the Southern Europe (Catalonia). *PhD Thesis*. Universitat Autònoma de Barcelona.

Bowker, L., & Pearson, J. (2002). *Working wiht specialized language: a practicle guide to using corpora.* London: Routledge.

Brickley, D., & Miller, L. (2014). *FOAF Vocabulary Specification 0.99.* Retrieved 05 2020, from Namespace Document 14 January 2014 - Paddington Edition: http://xmlns.com/foaf/spec/#term_publications

Cassidy, J. R. (1967). Aristotle on Definitions. *SOUTHERN JOURNAL OF PHILOSOPHY*, 110-118.

Chatterjee, S., & Nath, A. (2017). Auto-Explore the Web – Web Crawler. *International Journal of Innovative Research in Computer and Communication Engineering*, 6607-6618. Retrieved 7 2020, from https://www.researchgate.net/publication/316601171_Auto-Explore_the_Web_-_Web_Crawler

CIPR V5. (2006). Código Internacional das Práticas Rolheiras. (5ª versão). Santa Maria da Feira / Lisboa: C.E.Liège.

Condamines, A. (2005). Linguistique de corpus et terminologie. In L. Depecker (Ed.), *La terminologie: nature et enjeux* (pp. 36-47). Paris: Armand Colin.

Condamines, A., & Rebeyrolle, J. (2001). Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB): Method and Results. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology* (Vol. 2, pp. 127-148). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Conrad, S. (2011). Variation in corpora and its pedagogical implications. In E. Tognini-Bonelli, & W. Teubert (Eds.), *Perspectives on Corpus Linguistics* (Vol. 48, pp. 47-62). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Costa, A., & Pereira, H. (2004). Caracterização e Análise de Rendimento da Operação de Traçamento na Preparação de Pranchas de Cortiça para a Produção de Rolhas. *Silva e Lusitana, 12*(1), 51 - 66. Retrieved from http://repositorio.filcork.pt/community/#all

Costa, R. (2001). Pressupostos teóricos e metodológicos para a extracção automática de unidades terminológicas multilexémicas. *PhD Thesis*. Lisboa: Universidade Nova de Lisboa; Faculdade de Ciências Sociais e Humanas.

Costa, R. (2006). Plurality of Theoretical Approaches to Terminology. In P. Heribert (Ed.), *Linguistic Insights. Studies in Language and Communication* (Vol. 36). Berlin - Bern: Peter Lang Verlag.

Costa, R. (2013). Terminology and Specialised Lexicography: two complementary domains.

Costa, R. (2017). Les normes en terminologie. Que faire des synonymes? *Normes linguistiques et terminologiques : conflits d'usages*(110), pp. 45-57. doi:10.15122/isbn.978-2-406-07057-3.p.0045

Costa, R., & Silva, R. (2008). De la typologie à l'ontologie de textes. *Terminologies & Ontologies : Théories et applications(Actes de la 2ème Conférence - Toth Annecy - 2008)*. Annecy: Institut Porphyre.

Costa, R., Silva, R., Barros, S., & Lucas Soares, A. (2012). Mediation strategies between terminologists and experts.

Cruse, A. (2002). Hyponymy and Its Varieties. In R. Green (Ed.), *The semantics of relationships* (pp. 3-21). Springer Science+Business Media Dordrecht.

De Bessé, B. (1990). La définition terminologique. *Actes du Colloque la Définition, organisé par le CELEX (Centre d´études du Lexique) de l'Université Paris-Nord (1988)* (pp. 252 - 261). Paris: Larousse.

Depecker, L. (2015). How to build terminology science? In Kockaert, & Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 34-44). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., Marcellesi, J.-B., & Mével, J.-P. (2002). *Dictionnaire de linguistique.* Paris: Larousse.

EUFORGEN. (2020). *European Forest Genetic Resources Programme*. Retrieved 07 28, 2020, from Quercus Suber | Cork Oak: http://www.euforgen.org/species/quercus-suber/

EUROSTAT. (2019). *Agriculture, forestry and fishery statistics books.* Luxembourg: Publications Office of the European Union. doi:10.2785/798761

FAO. (2018). *State of Mediterranean Forests 2018.* Rome: the Food and Agriculture Oganization of the United Nations. Retrieved from https://planbleu.org/sites/default/files/publications/somf2018.pdf

Felber, H. (1984). *Terminology Manual.* Paris: Unesco and Inforterm.

Flowerdew, L. (2001). Corpus Linguistics in EST: a genre-based perspective. In A. Furness, G. Wong, & L. Wu (Eds.), *Penetrating Discourse: Integrating theory with practice* (pp. 21-40). Hong Kong : Language Centre, Hong Kong University of Science and Technology.

Geeraerts, D. (2015). *Handbook of Terminology* (Vol. 1). (H. Kockaert, & F. Steurs, Eds.) Amsterdam / Philadelphia: John Benjamins Publishing Company.

Gil, L. (1998). *Cortiça: produção, tecnologia e aplicação.* Lisboa: Instituto Nacional de Engenharia e Tecnologia Industrial.

Gil, L. (2002). *A Rolha de Cortiça e a sua Relação com o Vinho.* Portalegre: Agrupamento de Produtores Agrícolas e Florestais do Norte do Alentejo - APAFNA.

Gil, L. (2004). *Cortiça: da árvore aos produtos finais.* S. Brás de Alportel, Algarve: PELCOR.

Gil, L. (2007). Cork as a Building Material | Technical manual. Santa Maria de Lamas: APCOR - Portuguese Cork Association. Retrieved from https://www.apcor.pt/portfolio-posts/a-cortica-como-material-de-construcao-manual-tecnico/

Gil, L. (2014). Cork: a strategic material. (S. L. Suib, Ed.) *Frontiers in chemistry, 2*(16). Retrieved from www.frontiersin.org

Gil, L. (2015). Novas aplicações da cortiça. 36-38. INGENIUM. Retrieved from https://www.researchgate.net/publication/282973681_Novas_aplicacoes_da_cortica

Gil, L., & Varela, M. C. (2008). EUFORGEN Technical Guidelines for genetic conservation and use for cork oak (Quercus suber). Rome, Italy: Bioversity International. Retrieved from www.euforgen.com

Gries, S. (2011). Methodological and interdisciplinary stance in Corpus Linguistics. In E. Tognini-Bonell, & W. Teubert (Eds.), *Perspectives on Corpus Linguistics* (Vol. 48, pp. 81-98). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Halliday, M. A., Teubert, W., Yalop, C., & Cermáková, A. (2004). *Lexicology and corpus linguistics - An introduction.* London / NY: Continuum.

Halskov, J., & Barrière, C. (2010). Web-based extraction of semantic relation instances for terminology work. In A. Auger, & C. Barrière (Eds.), *Probing Semantic Relations:*

Exploration and identification in specialized texts (Vol. 23, pp. 20-42). Amsterdam | Philadelphia: John Benjamins B.V.

Horridge, M. (2011, March 24). A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools.

Horridge, M., Drummond, N., Goodwin, John, R. A., Stevens, R., & Wang, H. H. (2006). The Manchester OWL Syntax. *Proceedings of the OWLED*06 Workshop on OWL: Experiences and Directions, Athens, Georgia, USA*. Retrieved from https://dblp.uni-trier.de/db/conf/owled/owled2006.html

ICNF. (2019). *6º Inventário Nacional Florestal - 2015 Relatório final.* Instituto da Conservação da Natureza e das Florestas. Lisboa: Instituto da Conservação da Natureza e das Florestas. Retrieved from https://www.icnf.pt/noticias/inventarioflorestalnacional

INETI. (2001). *Guia Técnico Sectorial - Indústria da Cortiça.* Lisboa: Instituto Nacional de Engenharia e Tecnologia Industrial.

INPI. (2005, 12). A utilização e a valorização da propriedade industrial no sector da cortiça. *Colecção leituras e propriedade industrial, III*. Lisboa: INPI. Retrieved 04 10, 2014, from Instituto Nacional da Propriedade Industrial: http://www.marcasepatentes.pt/index.php?section=270

iO10551. (2017). *Indústrias de base florestal | Sinopse.* República Portuguesa, Direção - Geral das Atividades Económicas, Lisboa. Retrieved from https://www.dgae.gov.pt/servicos/politica-empresarial/setores-industriais/industrias-de-base-florestal.aspx

ISO 1087-1. (2000, 10 15). Travaux terminologiques — Vocabulaire —. *(Norme Internationale), Première édition*. Suisse: © ISO 2000.

ISO 704. (2009). Travail terminologique - Principes et méthodes. *NF ISO 704, 1er tirage 2009-12-P*. La Plaine Saint-Denis: Association Française de Normalisation.

ISO/FDIS 1087. (2019 (E)). Terminology work and terminology science - Vocabulary. Suisse: ISO.

Johansson, I. (2008). Four Kinds of Is_a Relation. In K. Munn, & B. Smith (Eds.), *Applied Ontology: An Introduction* (pp. 235-253). Frankfurt: Ontos Verlag.

Johansson, S. (2011). A multilingual outlook of corpora studies. In E. Tognini-Bonelli, & W. Teubert (Eds.), *Perspectives on corpus linguistics* (pp. 115-129). Amsterdam / Phildelphia: John Benjamins Publishing Company.

Kleiber, G. (1997). Sens, référence et existence : que faire de l'extra-linguistique ? In *Languages* (Langue, praxis et production de sens ed., Vol. 127, pp. 9-37). doi:10.3406/lgge.1997.2123

Kocourek, R. (1985). Terminologie et efficacité de la communication: critères linguistiques. *30*, 119-128. Montréal: Les Presses de l'Université de Montréal.

Lacasta, J., Nogueras-Iso, J., & Zarazaga-Soria, F. J. (2010). *A representation framework for terminological.* (R. Jain, & A. Sheth, Eds.) New York Dordrecht Heidelberg London: Springer Science+Business Media, LLC.

Lacy, W. L. (2005). *OWL: Representing Information Using the Web Ontology Language.* Crewe,Cheshire: Trafford Publishing (UK) Ltd.

Laviosa, S. (2011). Corpus Linguistics and translation studies. In E. Tognini-Bonelli, & W. Teubert (Eds.), *Perspectives on Corpus Linguistics* (pp. 131-153). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Leech, G. (2011). Principles and applications of Corpus Linguistics. In E. Tognini-Bonelli, & W. Teubert (Eds.), *Perspectives on Corpus Linguistics* (pp. 155-170). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Leech, G. (2013). Introducing corpus annotation. In R. Garside, G. Leech, & T. McEnery (Eds.), *Corpus Annotation: Linguistic Information From Computer Text Corpora* (pp. 1-18). New York: Routledge.

Lexical Computing. (2020). *Glossary*. Retrieved 07 2020, from Sketch Engine: https://www.sketchengine.eu/guide/glossary/

Lexical Computing. (2020). *Portuguese FreeLing part-of-speech tagset*. Retrieved 07 2020, from Sketch Engine: https://www.sketchengine.eu/portuguese-freeling-part-of-speech-tagset/

Lexical Computing. (2020). *Word sketch – collocations and word combinations*. Retrieved 07 05, 2020, from Sketch Engine: https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/

L'Homme, M. C. (2004). *La Terminologie: principes et techniques - Paramètres.* Montréal, Canadá: Les presses de l'Université de Montréal.

Lim, E., Liu, J., & Lee, R. (2011). *Knowledge Seeker – Ontology Modelling for Information Search and Management.* (J. Kacprzyk, Jain, & Lakhmi, Eds.) Hong Kong: Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-642-17916-7

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Massachussetts: Massachussetts Institute of Technology.

Marshamnn, E. (2007). Towards strategies for processing relationships between multiple relation participants in knowledge patterns. *Terminology*, pp. 1-34.

Marshamnn, E., L'Homme, M.-C., & Surtees, V. (N/A). Marqueurs de la relation cause-effect: stabilité et variation dans des corpus de nature différente.

Marshmann, E. (2003). Construction et gestion des corpus: Résumé et essai d'uniformisation du processus pour la terminologie.

McCarthy, M., & O'Keeffe, A. (2010). Historical perspective: what are corpora and how have they evolved? In M. McCarthy, & A. O'Keeffe (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 3-13). NY, USA and Canada: Routledge Handbooks.

McEnery, T. (2003). Corpus Linguistics. In R. Mitkov (Ed.), *Computational linguistics* (pp. 431-443). Oxford / NY: Oxford University Press.

McEnery, T., & Hardie, A. (2013). The History of Corpus Linguistics. In K. Allan (Ed.), *Oxford Handbooks in Linguistics.* Oxford University Press.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics - An introduction* (2 ed.). Edinburgh: Edinburgh University Press.

Meyer, I. (2001). Extracting Knowledge-Rich contexts for terminography: a conceptual and methodological framework. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology* (Vol. 2, pp. 279-302). Amsterdam / Philadelphia: John Benjamins B.V.

Minio-Paluello, L. (2016). *Categorias da Interpretação - Aristóteles* (Mesquita, António ed., Vols. I - Tomo II). (R. Santos, Trans.) Lisboa: Imprensa National - Casa da Moeda; Centro de Filosofia da Universidade de Lisboa.

Napoli, A. (1997). *Une introduction aux logiques de descriptions.* INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE. INRIA.

Nardi, D., & Brachmanv, R. J. (2003). An Introduction to Description Logics. In F. Baader, & D. Calvanese (Eds.), *The description logic handbook: theory, implementation, and applications* (pp. 1 - 40). NY, United States: Cambridge University Press.

Neveu, F. (2015). *Dictionnaire des sciences du langage* (2e édition revue et augmentée ed.). Paris: Armand Colin.

Norma Mínima V.1. (2007). Guia Internacional de Compra de Rolhas para Vinhos Tranquilos. *A trabalhar com o Comércio (trade) e com a Indústria da Cortiça*(Versão 1).

NP ISO 633. (2011). *Cortiça - vocabulário*. (IPQ) Retrieved Mai 29, 2014, from IPQ - Instituto normativo online: http://secure.ipq.pt/docs/AcervoOnline/INI_Pesq-AcervoOnlineFF.asp

Nunes, P. (2013). Análise do fluxo de processo industrial e do respetivo plano de inspeção e ensaios. *Ma dissert.* Porto: FEUP Universidade do Porto.

Pavel, & Nolet. (2002). *Manual de Terminologia.* (E. Faulstich, Trans.) Canadá: © Ministro de Obras Públicas e Serviços Governamentais do Canadá.

Pearson, J. (1998). *Terms in context.* Amsterdam: John Benjamins B.V.

Pereda, I. (2008, Maio). *Joaquim Vieira Natividade 1899-1968, Ciência e política do sobreiro e da cortiça.* Retrieved 08 28, 2014, from Pluridoc: www.pluridoc.com

Pereira, H. (2007). *Cork: Biology, Production and Uses.* Amsterdam: Elsevier.

Pomikálek, J. (2011). Removing Boilerplate and DuplicateContent from Web Corpora. *Ph.D. Thesis*. Brno. Retrieved 07 2020, from https://is.muni.cz/th/45523/fi_d/phdthesis.pdf

Pottier, B. (1992). *Théorie et analyse en Linguitique* (2, corrigée ed.). Paris: HACHETTE, Supérieur.

Pribbenow, S. (2002). Meronymic Relationships: From Classical Mereology to Complex Part-Whole Relations. In e. a. Rebecca Green (Ed.), *The semantics of relationships: An Interdisciplinary Perspective* (Vol. 3, pp. 35-48). Amsterdam: Springer-Science+Business Media, B.V.

Protégé. (2020). *Class Expression Syntax.* Retrieved 05 2020, from Protégé 5 Documentation: http://protegeproject.github.io/protege/class-expression-syntax/

Ramos, M. (2015, Setembro). O valor das definições para a organização conceptual da rolha da cortiça: uma questão de terminologia. *Thesis (Mast.)*. Lisboa: Faculdade de Ciências Sociais e Humanas - Universidade NOVA de Lisboa.

Ramos, M., Costa, R., & Roche, C. (2019). Dealing with specialised co-text in text mining: Verbal terminological collocations. *Terminology and Text Mining | Conference TOTh 2019.* Chambéry: Université Savoie Mont Blanc.

Rector, A. L. (2003). Modularisation of Domain Ontologies Implemented in Description Logics and related formalisms including OWL. *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, pp. 121-128. doi:https://doi.org/10.1145/945645.945664

Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., . . . Wroe, C. (2004, October). OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns. *Engineering Knowledge in the Age of the Semantic Web, 14th International ; EKAW 2004, Whittlebury Hall, UK, October 5-8, 2004, Proceedings*. doi: 10.1007/978-3-540-30202-5_5

Rey, A. (1979). *Que sais-je? La terminologie: noms et notions* (1re édition ed.). Paris: Presses Universitaires de France.

Rey, A. (1990). Définir la définition. In J. Chaurant, & F. Mazière (Ed.), *Actes du Colloque la Définition, organisé par le CELEX (Centre d´études du Lexique) de l'Université Paris-Nord (1988)* (pp. 13-22). Paris: Librairie Larousse.

Rey-Debove, J. (1998). *La linguistique du signe: Une approche sémiotique du langage.* Paris: Armand Colin.

Roche, C. (2005). Terminologie et ontologie. (L. Depecker, Ed.) *Langages*(157). doi:https://doi.org/10.3406/lgge.2005.974

Roche, C. (2007). Le terme et le concept : fondements d'une ontoterminologie. *TOTh 2007 | Terminologie & Ontologie : Théories et Applications.* Annecy.

Roche, C. (2009). La modélisation des concepts. *Dans tous les sens du terme*. (L. D.-J. Jean Quirion, Ed.) Ottawa: Les presses de l'Université d'Ottawa.

Roche, C. (2012). Should Terminology Principles be re-examined? pp. 17-32. Retrieved from https://hal.archives-ouvertes.fr/hal-01180279

Roche, C. (2015). Ontological definition. In F. Steurs, & H. Kockaert (Eds.), *Handbook of Terminology* (pp. 128-152). Amsterdam / Philadelphia: John Benjamins.

Rosen, K. H. (2000). *Handbook of discrete and combinatorial mathematics.* (J. MICHAELS, J. L. GROSS, J. W. GROSSMAN, & D. R. SHIER, Eds.) Boca Raton; London; New York; Washington D.C.: CRC Press.

Sager, J. C. (1990). *A practical course in terminology processing.* Amsterdam/Philadelphia: John Benjamins.

Santos, C., & Costa, R. (2015). Semasiological and onomasiological knowledge Representation. In H. J. Kockaert, & F. Steurs (Eds.), *Handbook of terminology* (Vol. 1, pp. 153-179). Amsterdam / Philadelphia: John Benjamins B.V.

Sardinha, T. B. (2011). Corpus Linguistics in South America. In E. Tognini-Bonelli, & W. Teubert (Eds.), *Perspectives on Corpus Linguistics* (Vol. 48, pp. 29-45). Amsterdam / Philadelphia: John Benjamins Publishing Company.

Sinclair, J. (1991). *Corpus, Concordance and Collocation.* Oxford: Oxford University Press.

Sinclair, J. (2004). *Trust the text - Language, corpus and discourse.* (R. Carter, Ed.) London and NY: Routledge.

Sinclair, J., & Ball, J. (1996). Preliminary Recommendations on Text Typology. EAGLES.

Smart, B. H. (1849). *Manual of Logic.* London: Longman, Brown, Green and Longmans.

Smith, R. (2020). *Aristotle's Logic*, Summer 2020 Edition. (E. N. Zalta, Editor, S. U. Center for the Study of Language and Information (CSLI), Producer, & Library of Congress Catalog Data:) Retrieved 04 27, 2020, from The Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/archives/sum2020/entries/aristotle-logic/

Spies, M., & Roche, C. (2006). Aristotelian ontologies and OWL modeling. In I. Johansson, T. Roth-Berghofer, & B. K.-B. Ingvar Johansson (Ed.), *WSPI 2006: Contributions to the Third International Workshop on Philosophy and Informatics.* Saarbrücken and Kaiserslautern: Universitat des Saarlandes. doi:10.4018/978-1-59904-660-0.ch002

StanfordEdu. (2020). *OntoGraf by Sean Falconer.* Retrieved from ProtégéWiki: https://protegewiki.stanford.edu/wiki/OntoGraf

Taber, G. (2009). *To cork or not to cork.* New York: SCRIBNER.

Teubert, W., & Cermáková, A. (2004). Directions in Corpus Linguistics. In Halliday, Teubert, Yallop, & Cermáková, *Lexicology and Corpus Linguistics* (pp. 113-165). London / New York: continuum.

Thoiron, P., & Béjoint, H. (2010, mars). *META: Translators' Journal, 55*(1), 105-118. doi:10.7202/039605ar

Tognini-Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 14-27). London / New York: Routledge.

Viana, V. (2011). The politics of Corpus Linguistics. In E. Tognini-Bonelli, & W. Teubert (Eds.), *Perspectives on Corpus Linguistics* (pp. 229-245). Amsterdam / Philadelphia: John Benjamins Publishing Company.

W3C. (1997). *Naming and Addressing: URIs, URLs, ...* Retrieved 05 2020, from W3C Architecture domain: https://www.w3.org/Addressing/

W3C. (2003). *OWL Web Ontology Language | W3C Candidate Recommendation 18 August 2003.* Retrieved from W3C Recommendation : https://www.w3.org/TR/2004/REC-owl-guide-20040210/#owl_AnnotationProperty

W3C. (2004). *OWL Web Ontology Language | Guide.* Retrieved from W3C Recommendation : https://www.w3.org/TR/2004/REC-owl-guide-20040210/

W3C. (2004). *OWL Web Ontology Language | Overview.* Retrieved from W3C Recommendation: https://www.w3.org/TR/owl-features/

W3C. (2004). *OWL Web Ontology Language Reference.* (M. Dean, & G. Schreiber, Eds.) Retrieved from W3C Recommendation 10 February 2004: https://www.w3.org/TR/owl-ref/

W3C. (2004). *OWL Web Ontology Language Semantics and Abstract Syntax.* Retrieved 05 2020, from W3C Recommendation: https://www.w3.org/TR/2004/REC-owl-semantics-20040210/syntax.html#2.3

W3C. (2005). *SKOS Core Guide | W3C Working Draft 2 November 2005.* Retrieved 04 2020, from W3C Recommendation: https://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/

W3C. (2012). *OWL 2 Web Ontology Language Manchester Syntax (Second Edition) W3C Working Group Note 11 December 2012.* Retrieved 2020, from W3C Recommendation: https://www.w3.org/TR/owl2-manchester-syntax/

W3C. (2014). *RDF 1.1 Concepts and Abstract Syntax - W3C Recommendation 25 February 2014.* Retrieved 04 22, 2020, from W3C Recommendation: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#section-Introduction

W3C. (2020). *OWL Web Ontology Language Reference [Enumerated Class].* Retrieved from W3C Recommendation: https://www.w3.org/TR/owl-ref/#EnumeratedClass

Williams, G. (2005). Introduction. In *La linguistique de corpus* (pp. 13-18). Rennes: Presses Universitaires de Rennes.

Winston, M., Chaffin, R., & Hermann, D. (1987, January). A Taxonomy of Part-Whole Relationships. *Cognitive Science*. Retrieved from https://www.researchgate.net/publication/245104866_A_Taxonomy_of_Part-Whole_Relationships

Wüster, E. (1998). *Introducción a la teoría general de la terminología y la lexicografía terminológica.* (M. T. CABRÉ, Ed., & A.-C. Nokerman, Trans.) Barcelona: Institut universitari de lingüística aplicada, Universitat Pompeu Fabra.

LIST OF FIGURES

LIST OF TABLES

ANNEXES

Annex 1

Cork, an ancient raw material

More pieces of evidence of the use of cork stoppers by Romans were found in shipwrecks from the 5th century BC. It is thought that Romans first used cork stoppers to protect the wine from air. Some amphoras dating from 500 BC have pieces of cork that were used to seal them; however, these cork stoppers and the ways in which they were used were quite different from the usages they currently have. Back then, a cork stopper was a large piece of raw bark that was fit into the mouth of the amphora and fixed in place with resin (see Taber). Romans are also known to have recommended the application of cork for beehives given its thermal properties, around the 2nd century BC, (see APCOR, 2019).

Taber (2009) further mentions that it was in Thebes, in a fresco found in a tomb from 1400 BC, that evidence was finally found of flat cylindrical tops on amphoras containing wine. As one can perceive, wine and cork have complemented each other since ancient times. This has also been proven by an amphora dating from the 1st century BC found in Ephesus: not only was it sealed with a cork stopper, but it also contained wine (see APCOR, 2019).

By circa 800 BC, a viticulture innovation had taken place, at the beginning of the rise of Greek city-states: the Greeks started to use three resins to seal wine containers. One of those resins was obtained from a terebinth tree – a species from the cashew family. The grape cultivation rapidly spread to Italy given the rise of Rome. It was Pliny, the Elder, whose writings date back to the 1st century of the modern era, that gave credit to the Celtic tribes from the Alpine valleys[190] for the introduction of wooden barrels to transport liquids instead of amphoras, mentioning the use of ***cork stoppers*** to keep the air out of the wooden casks. Pliny also describes in detail the procedure of harvesting a cork oak tree, and mentions that bark was mainly used for ships, namely "for anchor drag-ropes and fisherman's drag-nets and for the bung of casks, and also to make soles for women's winter shoes" (Taber, 2009, p. 10).

[190] Nowadays known as Switzerland (*ibid.*).

Furthermore, Pliny makes a novel and extensive reference to the cork oak in his famous Natural History. He explains that cork was worshipped in Greece as a symbol of freedom and honour, which is why only priests could cut it. In that same work, we read that the cork oak was consecrated by the Romans to Jupiter and that its leaves and branches served to crown the winning athletes (see APCOR, 2019). Furthermore, in the 2nd century AD, the Greek doctor Dioscorides pointed out some medicinal applications for cork – namely to treat hair loss (*ibid.*).

With the fall of the Roman Empire, in the 5th century – which coincided with the beginning of the Dark Ages in Europe – cork stoppers fell in disuse. Between 500 and 1500, trade decreased; thus, the Iberian Peninsula cork farmers could not sell their products to their neighbouring countries in the European continent. In addition to that reduced trade, the rise of the Moors in the Iberian Peninsula led to the prohibition of using cork with wine, since drinking alcoholic beverages was forbidden by their holy book. It was only in the 8th century that the conquering of Europe began – in the Battle of Tours, on 10 October 732, the Francs defeated the Moors 130 miles southwest from Paris. However, "it would take another century before the invaders were expelled from Iberia and a wine culture – and cork – returned to the area." (Taber, 2009).

According to Taber (2009), cork was first exported to England in 1307. This material was mainly used for soles in footwear manufacturing. It was later, in the 1500s, that cork was used in the form of a stopper to seal bottles holding wine. This author also argues that the "serendipitous union of corks and inexpensive glass bottles took place first in England and then spread to the Continent in the Seventeenth century" (p.12). Furthermore, as a piece of curiosity, the word "cork" (as a stopper in English) is said to have been used for the first time in one of Shakespeare's plays – *As you like it* (c.1600) (*ibid.*).

Also pointed out by Taber, there is a drawing of a man carving a cork stopper into shape to fit the bottle's neck in the well-known first encyclopaedia of Denis Diderot in 1751. In this encyclopaedia, there is "a detailed description of cork making. It said corks were used for shoes and slippers "but above all to close jugs and bottles"" (p.13).

Given our penchant for lexicography, one of our favourite historical facts dates back to the 17th century, in England, where the physicist Robert Hooke (1635 - 1703) was able to obtain the first microscopic image of cork using a microscope – a revolutionary device he had developed himself.

In his work, namely *Micrographia: Or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses with Observations and Inquiries Thereupon* [191] (1665), Hooke describes and draws his observations after cutting a thin piece of cork and observing its microscopic structure for the first time, in the scientific history of cork. According to his writings, Hook could perceive that cork was a solid substance with pores organised like a honeycomb structure. Furthermore, Hooke advanced the following reflection, where an explanation for the floating property is put forth:

> our *Microscope* informs us that the substance of Cork is altogether fill'd with Air, and that Air is perfectly enclosed in little Boxes or Cells distinct from one another. It seems very plain, why neither the Water, nor any other Air can easily insinuate itself into them, since there is already within them an *intus existens,* and consequently, why the pieces of Cork become so good floats for Nets, and stopples for Viols, or other close Vessels. (Hooke, p.113)

We can understand after Hooke's words that beyond floating devices and sealing devices, cork was also used as parts of musical instruments, for its anti-vibration properties had already been discovered.

After Hooke's observations, the word "cell" was coined, for Hooke had discovered plant cells through the observation of the cell walls in cork tissue and the small-box-like cells of cork reminded him of the cells of a monastery. A discovery that stands as the building blocks of all living things (see Taber, 2009). These observations were gracefully drawn, as shown below.

[191] "Hooke's reputation in the history of biology largely rests on his book Micrographia, published in 1665. Hooke devised the compound microscope and illumination system shown above, one of the best such microscopes of his time, and used it in his demonstrations at the Royal Society's meetings. With it he observed organisms as diverse as insects, sponges, bryozoans, foraminifera, and bird feathers. Micrographia was an accurate and detailed record of his observations, illustrated with magnificent drawings" https://ucmp.berkeley.edu/history/hooke.html .

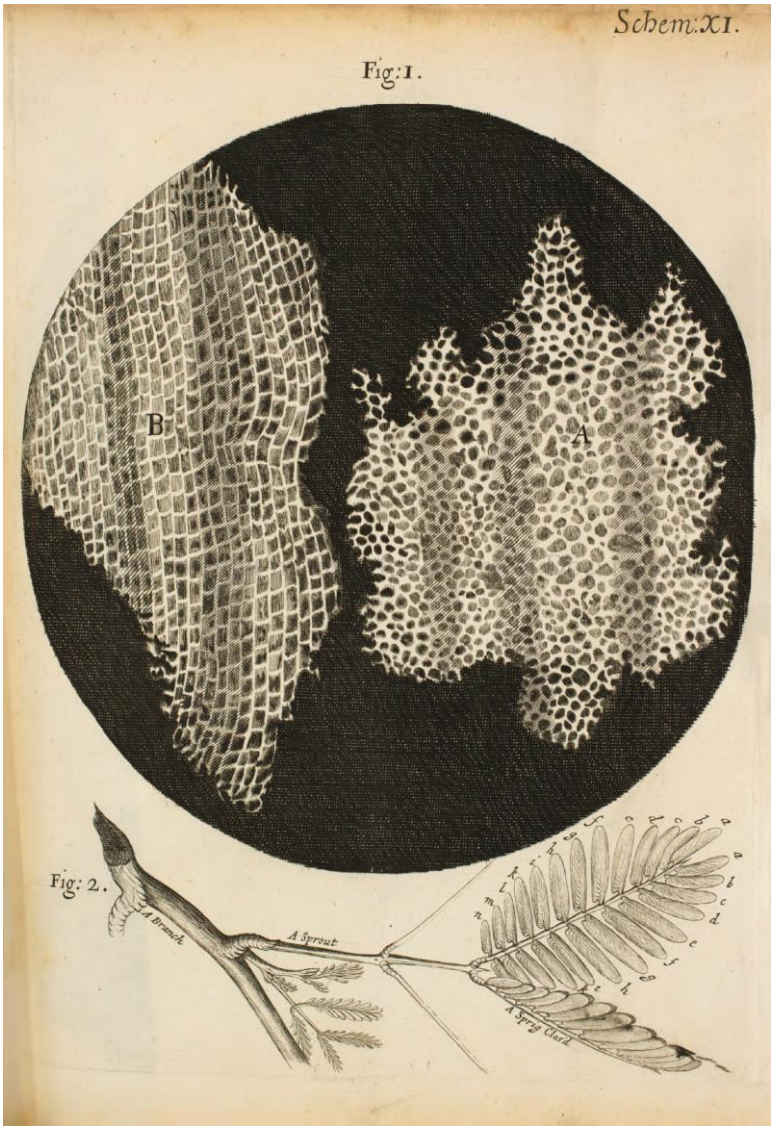


Figure 66: Hooke's drawing of his observations after a piece of cork observed by microscope[192]

The emergence of new instruments and applications – From the 19th century up to the current days

The following diagram is a compact and summarised representation of the cork evolution regarding the emergence of new technologies and applications, from the 19th century until the current days. It is not meant to be exhaustive but merely to represent a chronological evolution, in which we highlight the most relevant facts. For its elaboration, we resorted to data disclosed by APCOR[193] – the Portuguese Association for Cork.

[192] Document publicly disclosed by the Hunt Institute for Botanical Documentation - Hunt Library Carnegie Mellon University, available at http://www.huntbotanical.org/admin/uploads/hibd-hooke-micrographia-plates.pdf.

[193] https://www.apcor.pt/media-center/publicacoes/

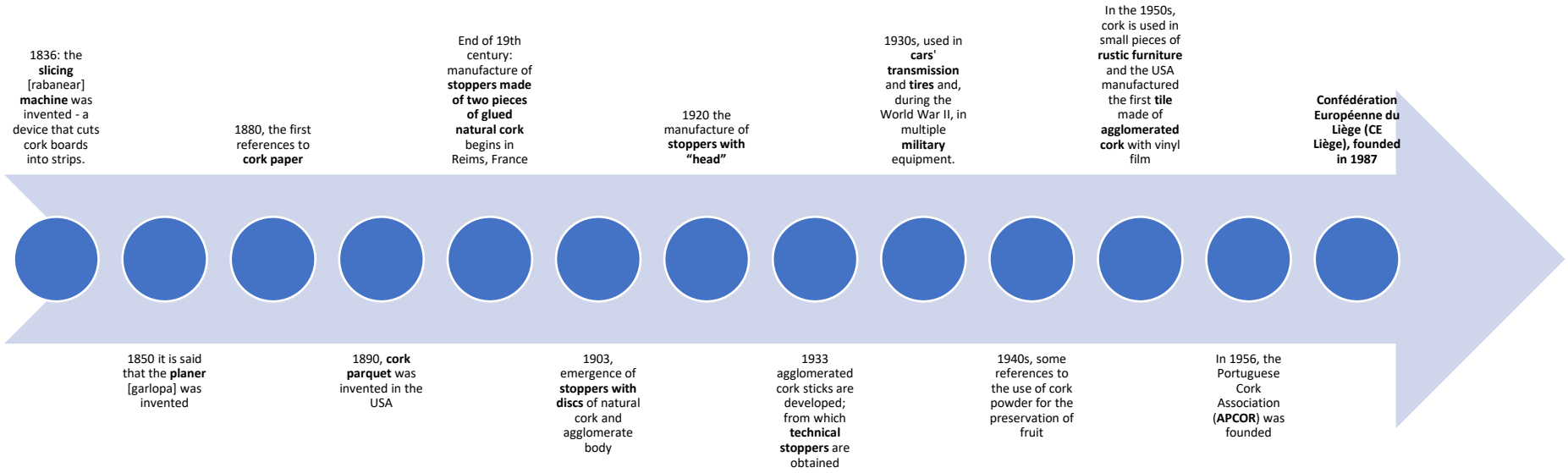
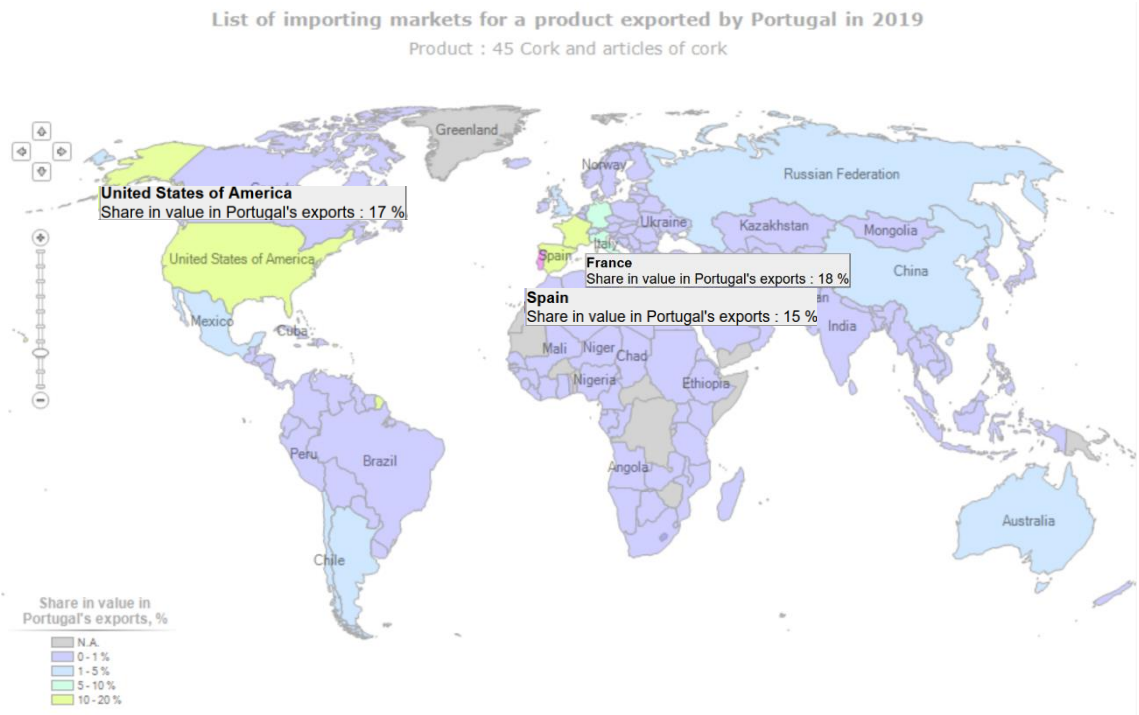
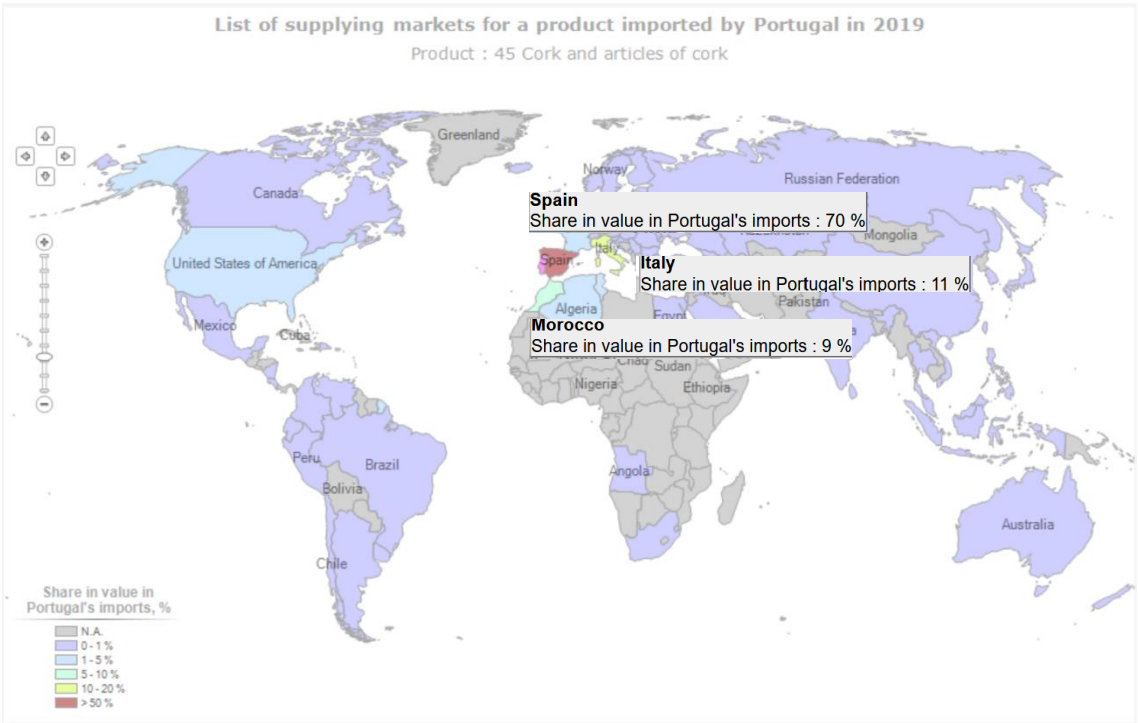


Figure 67: Chronological evolution of cork and instruments for its transformation, in which we highlight the most relevant facts from the 19th century up to the current days, based on APCOR – História da Cortiça

Annex 2



SHARE IN VALUE IN PORTUGAL'S **EXPORTS**. SOURCE: ITC CALCULATIONS BASED ON UN COMTRADE STATISTICS.



SHARE IN VALUE IN PORTUGAL'S **IMPORTS**. SOURCE: ITC CALCULATIONS BASED ON UN COMTRADE STATISTICS.

Annex 3

Organisation of the three subsectors of cork and corresponding activities, based on Gil (1998); Bicho (2004); INETI (2001); INPI (2005).

Cork subsector	Activity	Operation	Raw material
1st subsector	**Production**		
		planting	
		maintenance	
		debarking	
		selection	
			of cork
	Preparation	(activity binding the forest production and the commercialisation of the raw material)	
	by	boiling	
		slicing	
		trimming	
		choosing	
		baling	
			of the cork Amadia (reproduction cork)
2nd subsector	**Transformation**	(activity associated with the activity of Preparation)	
	by	simple carving	
		cutting	
			of the cork Amadia (reproduction cork)
3rd subsector	by	Granulating/Milling	
			of wastes **Resulting from the operation**
			trimmings
			shreds
			chunks
			of punching
			lowering
			chamfering
			other
		Regranulation	
			of granulated wastes
			from milling
			of Virgin cork
			chocks and parings
			refused cork
			wastes from other cork processing operations
			trimmings
			cork stoppers with defects
			wastes of agglomerates
			other

Annex 4

Concordance obtained after CQL 7 (Page 1)

terminological analysis

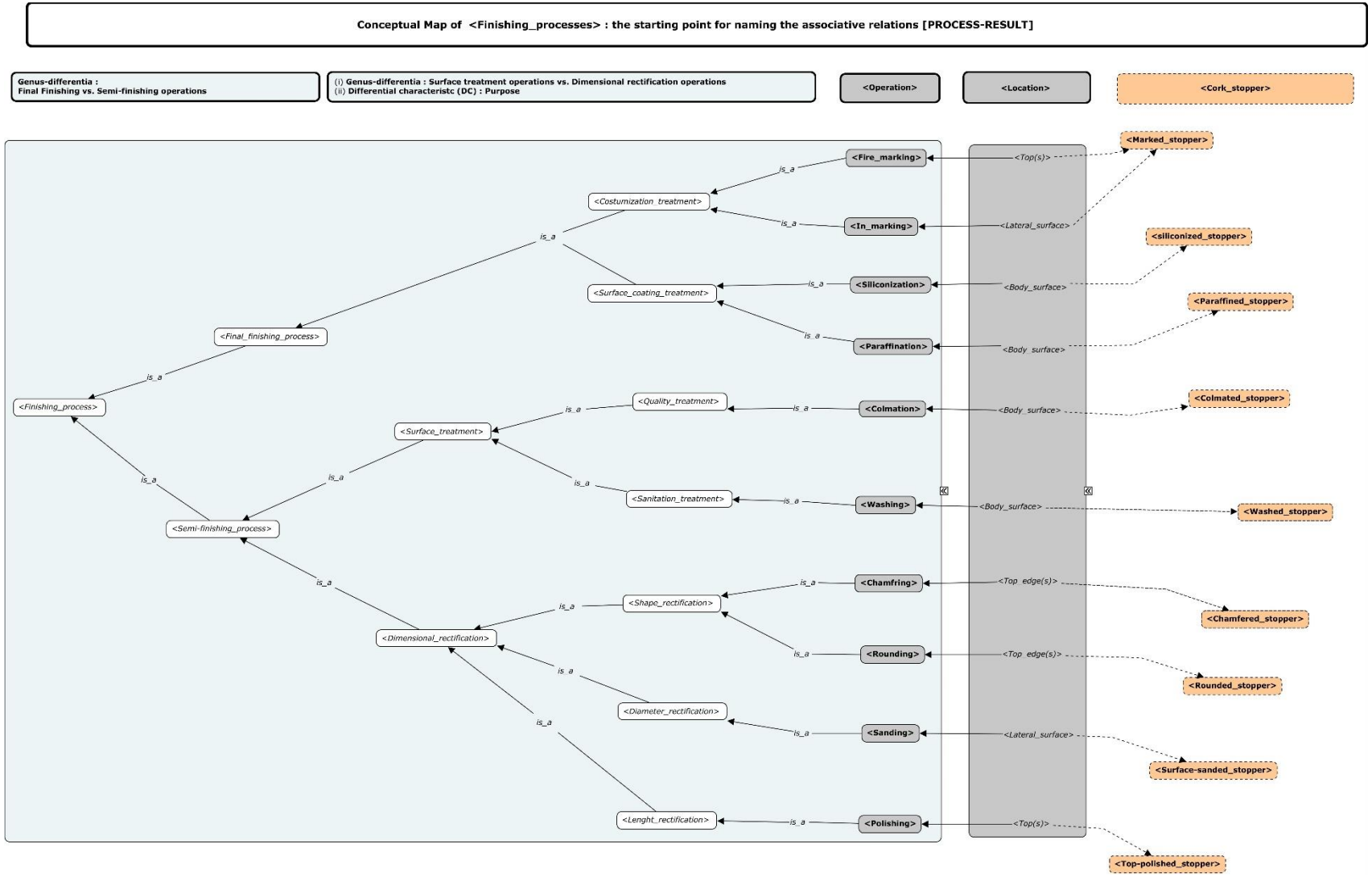
cql "rolha"[(tag="D.*")|(tag="S.*")]?[tag="A.*"]?"cortiça"?[]{… **167** (164.64 per million)

	Details	Left context	KWIC	Right context
101	doc#70	:/s><s> 876 <g/> . </s><s> A	**rolha de cortiça aglomerada**	é obtida pela mistura de granu
102	doc#70	:/s><s> 876 <g/> . </s><s> A	**rolha de cortiça aglomerada é obtida**	pela mistura de granulados co
103	doc#70	:/s><s> 877 <g/> . </s><s> A	**rolha de cortiça aglomerada**	obtida pela aglutinação de grâ
104	doc#70	:/s><s> 877 <g/> . </s><s> A	**rolha de cortiça aglomerada obtida**	pela aglutinação de grânulos d
105	doc#70	é vulgarmente designada por	**rolha de cortiça aglomerada**	nova geração ou rolha de micr
106	doc#70	:/s><s> 919 <g/> . </s><s> A	**rolha técnica para vinhos espumantes é produzida**	a partir de um corpo formado p
107	doc#70	s> Rolha de cortiça natural –	**rolha de cortiça feita**	totalmente de cortiça natural <
108	doc#70	le cortiça natural colmatada –	**rolha feita**	de cortiça natural em que são
109	doc#70	a aglomerada nova geração –	**rolha obtida**	pela aglutinação de grânulos d
110	doc#70	Jm ou Rolha Técnica <g/>) –	**rolha com um corpo de cortiça aglomerada**	e n discos de cortiça natural co
111	doc#78	a superior ao comprimento da	**rolha pretendida**	<g/> . </s><s> E Formação da
112	doc#78	a superior ao comprimento da	**rolha pretendida**	(<g/> NP 273 <g/>) <g/> . </
113	doc#78	ı/> . </s><s> O preço de uma	**rolha dita**	de qualidade extra ou superior
114	doc#78	selecção <g/> , em que cada	**rolha é avaliada**	indivi <g/> - dualmente e comp
115	doc#78	ais não devem abrir quando a	**rolha é dobrada**	pelos topos <g/> ; • ausência d
116	doc#78	zontais podem abrir quando a	**rolha é dobrada**	pelos topos <g/> ; • aceitável a
117	doc#79	seio , o que permite que uma	**rolha possa facilmente ser comprimida**	(<g/> para ser totalmente inse
118	doc#79	ma a optar pela dimensão da	**rolha mais ajustada**	ao cumprimento da sua função
119	doc#79	ma a optar pela dimensão da	**rolha mais ajustada**	ao cumprimento da sua função
120	doc#79	ma a optar pela dimensão da	**rolha mais ajustada**	ao cumprimento da sua função
121	doc#79	ma a optar pela dimensão da	**rolha mais ajustada**	ao cumprimento da sua função
122	doc#79	› 05.8 - Rolhas Capsuladas A	**rolha capsulada**	é uma rolha de cortiça em cujo
123	doc#79	das A rolha capsulada é uma	**rolha de cortiça em cujo topo é colocada**	uma cápsula <g/> , de madeira
124	doc#79	>) mais comuns são <g/> : A	**rolha capsulada**	é geralmente utilizada em vinh
125	doc#79	>) mais comuns são <g/> : A	**rolha capsulada é geralmente utilizada**	em vinhos licorosos <g/> / ger
126	doc#79	selecção do comprimento da	**rolha <g/> , deve ser calibrada**	para permitir um espaço de <g
127	doc#79	mente raro de aparecer numa	**rolha terminada**	<g/> ; • Defeitos de fabrico <g/
128	doc#79	itar uma torção exagerada da	**rolha <g/> . </s><s> Ao ser expelida**	<g/> , a rolha emitirá um som i
129	doc#84	neficios de utilização de uma	**rolha certificada**	<g/> , poderão obter certificaçã
130	doc#85	das como " <g/> 1+1 <g/> " –	**rolha constituída**	por um disco de cortiça natura
131	doc#89	› S Grossa > <g/> 18 linhas A	**rolha " <g/> 1+1 <g/> " é composta**	por um corpo de cortiça aglom
132	doc#91	damento de um dos topos da	**rolha <g/> , podendo ser utilizada**	para outros fins (<g/> especia
133	doc#91	um chanfro num dos topos da	**rolha <g/> , sendo realizada**	por lixagem <g/> . </s><s> • E
134	doc#93	seio , o que permite que uma	**rolha possa facilmente ser comprimida**	(<g/> para ser totalmente inse

Concordance obtained after CQL 7 (Page 2)

	Left context	KWIC	Right context
135	doc#93ma a optar pela dimensão da	**rolha mais ajustada**	ao cumprimento da sua funçãc
136	doc#93ma a optar pela dimensão da	**rolha mais ajustada**	ao cumprimento da sua funçãc
137	doc#93ma a optar pela dimensão da	**rolha mais ajustada**	ao cumprimento da sua funçãc
138	doc#93ma a optar pela dimensão da	**rolha mais ajustada**	ao cumprimento da sua funçãc
139	doc#93> 05.8 - Rolhas Capsuladas A	**rolha capsulada**	é uma rolha de cortiça em cujc
140	doc#93idas A rolha capsulada é uma	**rolha de cortiça em cujo topo é colocada**	uma cápsula <g/> , de madeira
141	doc#93>) mais comuns são <g/> : A	**rolha capsulada**	é geralmente utilizada em vinh
142	doc#93>) mais comuns são <g/> : A	**rolha capsulada é geralmente utilizada**	em vinhos licorosos <g/> / gen
143	doc#93 selecção do comprimento da	**rolha <g/> , deve ser calibrada**	para permitir um espaço de <g
144	doc#93mente raro de aparecer numa	**rolha terminada**	<g/> ; • Defeitos de fabrico <g/
145	doc#93itar uma torção exagerada da	**rolha <g/> . </s><s> Ao ser expelida**	<g/> , a rolha emitirá um som i
146	doc#94das como " <g/> 1+1 <g/> " –	**rolha constituída**	por um disco de cortiça natural
147	doc#95ara regularizar a superfície da	**rolha <g/> . </s><s> Selecção Vulgarmente designada**	por escolha <g/> , é a operaçã
148	doc#95ation <g/>) <g/> . </s><s> A	**rolha já montada**	é sujeita à secagem <g/> , de
149	doc#95s categorias <g/> : • • • • • • • •	**rolha natural – peça única <g/> , extraída**	por brocagem de um traço de
150	doc#95natural coladas entre si <g/> ;	**rolha natural colmatada**	- rolhas naturais cujos poros es
151	doc#95olhas normais <g/> . </s><s>	**rolha aglomerada**	– rolhas com um corpo de cort
152	doc#95 de cortiça aglomerada <g/> ;	**rolha micro granulada**	– rolhas com um corpo de cort
153	doc#95 entre 0,25 mm e 8 mm <g/> ;	**rolha capsulada**	– rolha de cortiça natural em c
154	doc#95<g/> , o que permite que uma	**rolha possa facilmente ser comprimida**	(<g/> para ser totalmente inse
155	doc#96 1982 <g/>) <g/> . </s><s> A	**rolha é então vazada**	por brocagem manual <g/> , s
156	doc#96</s><s> Entende-se que uma	**rolha é de qualidade elevada**	quando apresenta como carac
157	doc#96ual <g/>) <g/> , porque cada	**rolha tem que ser avaliada**	individualmente <g/> , e é efec
158	doc#96iência <g/>) <g/> , blocos de	**rolha aglomerada**	(<g/> alta frequência <g/>) <
159	doc#96- alta frequência <g/>) <g/> ,	**rolha aglomerada**	por extrusão e por moldação <
160	doc#96na de champanhe ou parte da	**rolha de champanhe que não foi introduzida**	no gargalo <g/> . </s><s> Top
161	doc#97rra de disco <g/> . </s><s> A	**rolha é então vazada**	por brocagem manual <g/> , s
162	doc#97</s><s> Entende-se que uma	**rolha é de qualidade elevada**	quando apresenta como carac
163	doc#97o longitudinal na superfície da	**rolha provocada**	por se brocarem as rolhas mui
164	doc#97ual <g/>) <g/> , porque cada	**rolha tem que ser avaliada**	individualmente <g/> , e é efec
165	doc#97 consumidores associam uma	**rolha marcada**	com um bonito desenho a um
166	doc#97ísticos da superfície lateral da	**rolha <g/> , por vezes mesmo colorida**	<g/> , para uso em gargalos ni
167	doc#97anquilos <g/> . </s><s> Esta	**rolha é referida**	como tendo a vantagem de ter

Annex 5


Conceptual Map of <Finishing_processes> : the starting point for naming the associative relations [PROCESS-RESULT]
Genus-differentia :
Final Finishing vs. Semi-finishing operations
(i) Genus-differentia : Surface treatment operations vs. Dimensional rectification operations
(ii) Differential characteristc (DC) : Purpose
<Operation>
<Location>
<Cork_stopper>
<Finishing_process>
<Final_finishing_process>
<Semi-finishing_process>
<Surface_treatment>
<Dimensional_rectification>
<Costumization_treatment>
<Surface_coating_treatment>
<Quality_treatment>
<Sanitation_treatment>
<Shape_rectification>
<Diameter_rectification>
<Lenght_rectification>
is_a
<Fire_marking>
<In_marking>
<Siliconization>
<Paraffination>
<Colmation>
<Washing>
<Chamfring>
<Rounding>
<Sanding>
<Polishing>
<Top(s)>
<Lateral_surface>
<Body_surface>
<Top edge(s)>
<Marked_stopper>
<siliconized_stopper>
<Paraffined_stopper>
<Colmated_stopper>
<Washed_stopper>
<Chamfered_stopper>
<Rounded_stopper>
<Surface-sanded_stopper>
<Top-polished_stopper>